IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 19 December 2024, accepted 9 January 2025, date of publication 13 January 2025, date of current version 17 January 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3528733

RESEARCH ARTICLE

Abstractive Summarization of Historical Documents: A New Dataset and Novel Method Using a Domain-Specific Pretrained Model

KEERTHANA MURUGARAJ, SALIMA LAMSIYAH, AND CHRISTOPH SCHOMMER
Department of Computer Science, Faculty of Science, Technology and Medicine, University of Luxembourg, 1855 Luxembourg City, Luxembourg
Corresponding author: Keerthana Murugaraj (keerthana.murugaraj@uni.lu)

ABSTRACT Automatic Text Summarization (ATS) systems aim to generate concise summaries of documents while preserving their essential aspects using either extractive or abstractive approaches. Transformer-based ATS methods have achieved success in various domains; however, there is a lack of research in the historical domain. In this paper, we introduce HistBERTSum-Abs, a novel method for abstractive historical single-document summarization. A major challenge in this task is the lack of annotated datasets for historical text summarization. To address this issue, we create a new dataset using archived documents obtained from the Centre Virtuel de la Connaissance sur l'Europe group at the University of Luxembourg. Furthermore, we leverage the potential of HistBERT, a domain-specific bidirectional language model trained on the balanced Corpus of Historical American English, (https://www.english-corpora.org/coha/) to capture the semantics of the input documents. Specifically, our method adopts an encoder-decoder architecture, combining the pre-trained HistBERT encoder with a randomly initialized Transformer decoder. To address the mismatch between the pre-trained encoder and the non-pre-trained decoder, we employ a novel fine-tuning schedule that uses different optimizers for each component. Experimental results on our constructed dataset demonstrate that our HistBERTSum-Abs method outperforms recent state-of-the-art deep learning-based methods and achieves results comparable to state-of-the-art LLMs in zero-shot settings in terms of ROUGE-1, ROUGE-2, and ROUGE-L F1 scores. To the best of our knowledge, this is the first work on abstractive historical text summarization.

INDEX TERMS Historical text summarization, abstractive approach, pre-trained HistBERT encoder, large language models, transfer learning.

I. INTRODUCTION

In the last few decades, the field of digital humanities has witnessed an enormous effort to digitize historical documents, resulting in an unprecedented volume of machine-readable texts available in digital format. While this development represents a significant breakthrough in terms of preservation and accessibility, it also presents new opportunities for content mining. The main challenge is to develop effective technologies for searching, retrieving, and exploring information from this vast collection of historical documents, commonly referred to as the *big data of the past* [2]. This challenge highlights also the need for extensive research in historical text processing, which can efficiently provide relevant historical information to both historians and ordinary citizens with minimal effort. Therefore, automatic text summarization (ATS) can be one of the effective tools to address the issue of information overload in the historical domain. This can have significant implications for research in fields such as history, archaeology, and digital humanities [3].

Automatic Text Summarization (ATS) is a challenging research field in Natural Language Processing (NLP) that

The associate editor coordinating the review of this manuscript and approving it for publication was Zijian Zhang.

aims to automatically generate concise summaries from a document or a collection of documents while preserving their essential aspects [4]. ATS systems generally rely on two commonly used approaches: *extractive* summarization and *abstractive* summarization. *Extractive* methods involve identifying and extracting the most significant sentences directly from the source documents without any modification [5], [6], [7]. In contrast, *abstractive* methods require language generation capabilities to produce summaries containing novel words or phrases not present in the original documents [8], [9], [10]. Furthermore, ATS systems can be categorized based on the number of input documents (single-document or multi-document) or the purpose of summarization (generic or query-focused). For instance, *generic* summaries aim to cover all relevant information from a source document without considering specific user needs [11]. On the other hand, *query-focused* summaries are tailored to address specific user queries or information needs [12]. In our research, we focus on the summarization of *generic single documents* in the historical domain using an *abstractive* approach. Specifically, we introduce *HistBERTSum-Abs*, a novel abstractive method designed to generate concise and coherent summaries for single historical documents. The goal of this work is to provide substantial benefits for historians and other users by saving time and delivering relevant information. Additionally, we position historical text summarization as a challenging research area with significant potential to advance NLP and text mining in the historical domain.

The effectiveness of supervised text summarization methods is closely tied to the quality of the datasets used for training the models. Indeed, benchmark datasets play a crucial role in developing and evaluating text summarization models for historical documents. However, the current ATS datasets, such as CNN/DailyMail [8] and arXiv [13], are not suitable for historical texts due to various reasons. For instance, historical texts often focus on the documentation of significant events, political decisions, and societal developments, which are rarely represented in general news articles or scientific papers—the main sources for current ATS datasets. Moreover, the historical context in which these texts were written adds layers of meaning and interpretation that go beyond modern summaries' typical focus. Capturing the significance of past events and their relationships requires models to understand timelines, causality, and event-specific knowledge. Consequently, models trained on existing datasets may struggle to summarize historical texts effectively, as they fail to account for the underlying historical narratives and context. To address this issue, we created a historical dataset for single-document summarization by collecting archived documents from the Centre Virtuel de la Connaissance sur l'Europe (CVCE)[1] group at the University of Luxembourg. We carefully selected and pre-processed these documents to prepare the master dataset for our research.

Abstractive text summarization based on pre-trained language models (PLMs), such as BERTSum [14], BART [15], T5 [16], and Pegasus [17], has achieved significant success. These models are typically trained on large-scale datasets such as CNN/DailyMail [8], PubMed [13], and arXiv [13]. Building upon this success, we leverage the potential of HistBERT encoder [18], a domain-specific language model trained on the balanced Corpus of Historical American English,[2] to capture the contextual representations of the input historical documents. Additionally, we incorporate a sentence position embedding mechanism in our approach. This mechanism enables the model to capture the position information of sentences within the document, allowing it to understand the structural features of the text. By considering the order of sentences, our model gains a better understanding of the document's structure and can generate more coherent summaries. Furthermore, the proposed method HistBERTSum-Abs is based on an encoder-decoder architecture that combines the pre-trained HistBERT encoder with a randomly-initialized Transformer decoder [1]. This architecture is trained on the constructed dataset for abstractive historical single document summarization. To tackle the discrepancy between the pre-trained encoder and the non-pre-trained decoder, we adopt a novel fine-tuning schedule that employs distinct optimizers for each component. This approach helps to alleviate the mismatch and optimize the training process effectively [14]. To the best of our knowledge, our work is the first attempt to address the task of summarizing historical texts in English. Consequently, there are no directly comparable systems available. Instead, we implemented recent state-of-the-art summarization models and conducted comprehensive evaluations to demonstrate the effectiveness of the proposed method. Therefore, our work serves as a solid foundation and starting point for future research in the domain of historical text summarization.

To summarize, the key contributions of our work are as follows:

- We construct a high-quality, gold-standard text summarization dataset consisting of English historical documents summarized by language experts.
- We propose a novel abstractive method for historical single-document summarization that leverages a domain-aware language model, *HistBERT*, pre-trained on a large-scale historical corpus. Additionally, we incorporate a sentence position embedding mechanism to effectively capture sentence position information and enhance the model's understanding of the document's structural features.
- We evaluate the proposed method on our historical dataset against recent state-of-the-art deep learning-based abstractive summarization models using fine-tuning, as well as against recent LLMs in a zero-shot setting. Experimental results demonstrate that the proposed method achieves competitive performance

[1] https://www.cvce.eu/en

[2] https://www.english-corpora.org/coha/

in terms of ROUGE-1, ROUGE-2, and ROUGE-L F1 scores. Consequently, our method can serve as a solid baseline for future historical text summarization research.

The remainder of this paper is organized as follows: Section II provides a review of related work. Section III describes the proposed method in detail. Section IV presents the experimental results, including a discussion and interpretation of the model's performance. Finally, Section V concludes the paper and outlines potential directions for future research.

II. RELATED WORK

In this section, we will review research on historical NLP applications and neural abstractive text summarization methods. For a detailed review, readers may refer to [9], [19], [20], [21], [22], [23], [24], and [25] surveys, respectively.

A. HISTORICAL NLP APPLICATIONS

Historical texts refer to documents that provide knowledge and insights about the human past, often requiring careful interpretation. They present a sequence of events that occurred in different time periods, enabling researchers to uncover condensed chronological accounts of significant events within specific locations and timeframes [26]. In recent years, the fields of digital humanities and natural language processing for historical texts have received increasing attention. Several workshops have been organized to explore historical NLP applications, including the Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH) workshop series, which has been held since 2007.[3] Other notable events include the Computational Historical Linguistics workshops, organized alongside the NoDaLiDa conferences in 2013 and 2017 [27], [28], and the Language Resources and Technologies for Processing and Linking Historical Documents and Archives (LRT4HDA) workshop, which took place during the LREC conference in 2014.Additionally, the International Conference on Natural Language Processing for Digital Humanities focuses on applying NLP techniques to digital humanities research. Topics include any aspect of digital humanities with a strong emphasis on natural language processing or text generation. The field of computational linguistic analysis for historical texts has also been explored in several noteworthy Ph.D. theses [29], [30]. Indeed, most existing NLP studies on historical documents have focused on tasks such as spelling normalization [21], [31], machine translation [32], and sequence labeling tasks, including part-of-speech tagging [33] and named entity recognition [22], [34]. However, with the success of deep neural networks, new applications have emerged, including sentiment analysis [35], information retrieval [30], event extraction [36], [37], and text classification [38].

Nevertheless, only a limited number of works have been proposed for historical text summarization tasks. For instance, Gung and Kalita [39] have introduced a new method that leverages temporal information to enhance the extractive summarization of historical texts. The proposed method involves clustering sentences by timestamp and temporal similarity, assigning an importance score to each cluster, and using it as a weight in standard sentence ranking techniques. This temporal weighting yields consistent improvements over baseline systems. In the same context, Ghosh et al. [40] have proposed an extractive method for Bangla document summarization, which uses graph-based sentence scoring features along with both surface and corpus-level features such as named entities, numerical data, and title words. The proposed method outperforms all existing methods used in Bangla news summarization based on the standard summary evaluation method ROUGE [41].Similarly, Peng et al. [42] have introduced an abstractive method for summarizing German and Chinese historical news documents in their corresponding modern language. The proposed method uses a cross-lingual transfer learning technique, which can be trained even with no cross-lingual (historical to modern) parallel data. The experimental results using automatic and human evaluations have demonstrated the strengths of the proposed method over state-of-the-art baselines. Our recent work [43] introduces a novel method for extractive single-document summarization, emphasizing the critical role of historical domain knowledge in generating effective summaries.

Despite the growing interest in historical NLP applications, historical text summarization remains largely unexplored. Most previous research on text summarization has focused on domains such as news [8], biomedical [13], [44], and scientific texts [13]. Summarizing historical texts, however, presents unique challenges rooted in the nature of historical documents. These texts often serve as critical sources for understanding past events, decisions, and societal developments, requiring careful interpretation to extract their key insights. Additionally, the limited availability of annotated resources specifically designed for historical summarization has limited progress in this area. To address these challenges, we created a new dataset tailored for summarizing single historical documents, enabling the development and evaluation of summarization methods in this domain. Furthermore, we leverage a domain-aware pre-trained language model to effectively capture the semantics and contextual meaning of historical narratives. Our work represents the first attempt to summarize English historical texts and aims to encourage further research in this promising and impactful area.

B. NEURAL ABSTRACTIVE TEXT SUMMARIZATION

Neural abstractive text summarization methods are mainly based on sequence-to-sequence models, where an encoder takes a sequence of tokens from the source document $t = [t_1, t_2, \ldots, t_n]$ and maps it to a sequence of continuous

[3] https://aclanthology.org/venues/latech/

representations $z = [z_1, z_2, \ldots, z_n]$. Then, the decoder generates the target summary $y = [y_1, y_2, \ldots, y_m]$ one token at a time following an auto-regressive manner. This means that it models the conditional probability of generating each token of the summary given the sequence of tokens in the source document, which is formally defined as $p = (y_1, y_2, \ldots, y_m | t_1, t_3, \ldots, t_n)$. As already mentioned, most existing works on abstractive text summarization methods have focused on contemporary texts where [8], [45] were among the first works that applied sequence-to-sequence architecture based on recurrent neural networks to abstractive text summarization.

In recent years, pre-trained language models based on the Transformer architecture [1] have shown impressive performance in a wide range of natural language processing tasks [46], [47], [48]. Indeed, PLMs have been mainly employed to improve the performance of language understanding tasks. Recently, there have been great efforts to extend the application of pre-trained models to various natural language generation tasks, including abstractive text summarization [14], [15], [16], [17]. In this context, Liu and Lapata [14] have introduced the BERTSum - single-document summarization model that is based on the pre-trained BERT language model [46]. The authors proposed a general framework for both extractive and abstractive models, which uses a new document-level encoder to capture the document semantics and obtain the sentence-level document representation. Moreover, Google AI researchers [16] have experimented with transfer learning techniques for various natural language processing tasks. They introduced a unified framework called text-to-text transfer transformer (T5) that converts all text-based language into the text-to-text format. It is trained on various unsupervised and supervised tasks with a uniform architecture. Several tasks like machine translation, document summarization, classification, and regression are cast into one framework as input and trained in the model to generate target text for each task. In the same context, Lewis et al. [15] have introduced BART model, a denoising auto-encoder for pre-training sequence-to-sequence models. BART is used for several natural language generation tasks, and it achieved new state-of-the-art results on a range of abstractive dialogue, question-answering, and summarization applications. Recently, Zhang et al. [17] have developed a novel system PEGASUS, which pre-training with extracted gap sentences for abstractive summarization. The PEGASUS model represents the state-of-the-art for abstractive summarization methods. In contrast to previous models (i.e., BERTSum [14], T5 [16], and BART [15]) that were trained by masking fragments or small continuous text spans, PEGASUS is trained by masking multiple complete sentences rather than small text fragments and concatenating the gap sentences into pseudo-summary. The pseudo-summaries are used as labels for model training.

Researchers have also explored the use of pre-trained language models in other specific domains, such as the summarization of legal case judgments [49]. In this context, LegalSumm [50], an abstractive summarization model for legal rulings, leverages textual entailment to produce concise and coherent summaries. Building on this progress, domain-specific abstractive summarization models, such as Legal-Pegasus—based on Google Pegasus [17]—have been made available on Hugging Face.[4] Another notable model, Legal-LED,[5] utilizes the Longformer Encoder-Decoder (LED) architecture [51] to address the challenges of lengthy legal texts. Additionally, there has been growing interest in summarizing legal documents in low-resource languages. For instance, a recent advancement in abstractive summarization focuses on Portuguese legal documents, leveraging contrastive learning techniques to improve summarization quality [52]. Inspired by the proven success of these models in the legal domain, we leverage the potential of pre-trained language models in this paper to address the unique challenges of the historical domain.

In contrast to existing works, we introduce a novel method for abstractive historical single-document summarization. The proposed method is primarily based on the HistBERT model [18], a pre-trained language model designed for diachronic lexical-semantic analysis. HistBERT has been trained on a large corpus of historical English documents, COHA,[6] enabling it to effectively capture and analyze the evolving meanings of words across different time periods. By leveraging HistBERT, our aim is to enhance the quality of historical document summarization. HistBERT's ability to capture semantic changes over time provides valuable insights for generating informative and coherent summaries that preserve the essence of the original texts. Moreover, we employ a document-level encoder based on HistBERT, which encodes entire documents and generates sentence-level representations. This approach is particularly crucial for text summarization tasks, as it requires a comprehensive understanding of natural language that goes beyond the meaning of individual words and sentences. Furthermore, given the success of large language models (LLMs) in various NLP tasks, including text summarization [24], [25], [53], [54], we further evaluate the performance of recent LLMs on historical text summarization. Specifically, we leverage models such as ChatGPT-4o Mini,[7] Mistral-7B-Instruct,[8] and Llama 3.1 8B Instruct[9] in a zero-shot setting to assess their capabilities in this domain.

III. PROPOSED METHOD

We propose a generic supervised abstractive method for summarizing historical documents, named as HistBERTSum-Abs. The proposed method uses the HistBERT model [18] as an encoder to capture sentence-level embeddings. As illustrated in Figure 1, the input document $d = \{S_1, S_2, \ldots, S_n\}$

[4] https://huggingface.co/nsi319/legal-pegasus

[5] https://huggingface.co/nsi319/legal-led-base-16384

[6] https://www.english-corpora.org/coha/

[7] https://platform.openai.com/docs/models/gp#gpt-4o-mini

[8] unsloth/mistral-7b-instruct-v0.3.

[9] unsloth/Meta-Llama-3.1-8B-Instruct-bnb-4bit.

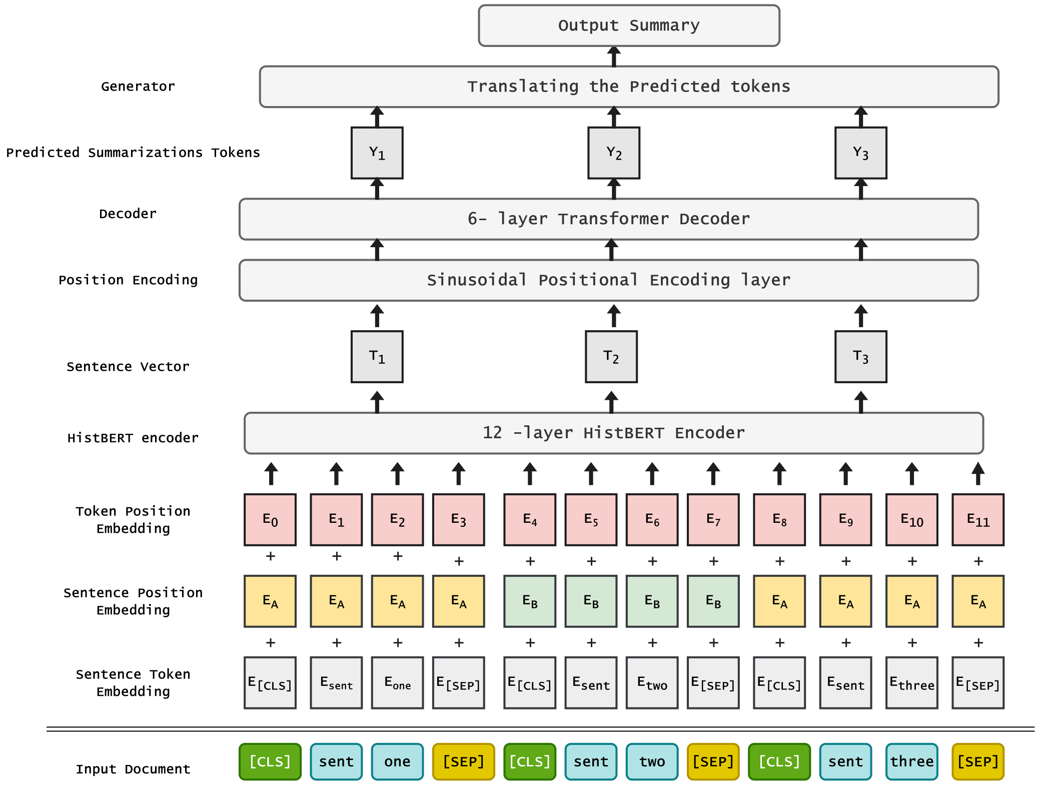
Output Summary
Generator
Translating the Predicted tokens
Predicted Summarizations Tokens
Y1
Y2
Y3
Decoder
6- layer Transformer Decoder
Position Encoding
Sinusoidal Positional Encoding layer
Sentence Vector
T1
T2
T3
HistBERT encoder
12 -layer HistBERT Encoder
Token Position Embedding
E0 E1 E2 E3 E4 E5 E6 E7 E8 E9 E10 E11
+
Sentence Position Embedding
EA EA EA EA EB EB EB EB EA EA EA EA
+
Sentence Token Embedding
E[CLS] Esent Eone E[SEP] E[CLS] Esent Etwo E[SEP] E[CLS] Esent Ethree E[SEP]
Input Document
[CLS] sent one [SEP] [CLS] sent two [SEP] [CLS] sent three [SEP]

FIGURE 1. HistBERTSum-Abs architecture. After adding two special tokens [CLS] and [SEP], the input document is passed to the three embedding layers. The final embedding representations are passed to the Encoder to obtain the sentence vectors. The decoder predicts the summarization tokens. The generator consists of linear and SoftMax layers that convert the decoder output tokens into words to form the final summary.

consists of n sentences where the sentence S_i is the $i-th$ sentence in the document, which consists of k tokens. The sentence embeddings vector is constructed by fusing token embeddings, sentence position embeddings, and token position embeddings. Then, the sentence embedding vector is fed into the HistBERT encoder to learn deep-level features of sentences based on a multi-head self-attention mechanism. Indeed, the HistBERT model uses the same architecture as BERT, but with parameters trained on a historical-domain corpus, and is employed as the encoder in our model due to its performance in various historical NLP tasks [18]. Finally, an encoder-decoder model, based on the HistBERT encoder and the transformer decoder [1], is fine-tuned on our dataset for the abstractive historical single-document summarization task. In the rest of this section, we will successively describe the different steps of our HistBERTSum-Abs method.

A. EMBEDDING REPRESENTATION

Embedding representation is an essential step for any natural language processing application. It involves encoding words or sentences into vectors that capture their semantics using, for instance, pre-trained language models (e.g., BERT, HistBERT). In particular, BERT [46] or HistBERT [18] models take a single sentence or a pair of sentences as input for downstream NLP tasks such as text classification. However, for text summarization, the input is an entire document that may contain multiple sentences. In our method, we use HistBERT encoder [18] to effectively capture the semantic meaning of the input historical document. As shown in Figure 1, we use three embedding layers, including sentence token embedding, sentence position embedding, and token position embedding. Each of these embedding layers serves a unique purpose in capturing different aspects of the input data, as described in the following:

1) SENTENCE TOKEN EMBEDDING

The sentence token embedding layer aims to encode the meaning of each token in the sentence. Formally, given a historical document d, we first use the Stanford CoreNLP toolkit [55] for sentence splitting. Thus, the document d is represented as a set of n sentences, denoted as $d = \{S_1, S_2, \ldots, S_n\}$, where each sentence S_i in d is then converted into tokens. In addition, we add two tokens **[CLS]** and **[SEP]** at the beginning and the end of each sentence, respectively. The **[CLS]** token is used to provide information about the sentence's features, while the **[SEP]** token helps the model understand the subsequent sentence. Noticing that the **[SEP]** token is also appended at the end of the document after the last sentence. The vocabulary constructed during HistBERT pretraining is used to index each word in the input document d. The sentence token embedding layer converts each token t_j in S_i into an embedded vector with fixed dimensions that is denoted as dim_{token}. The embedding vector that is obtained using the pre-trained HistBERT language model is formally defined in the following Equation:

$$EMB_{token}(t_j) = HistBERT(index(t_j)) \tag{1}$$

where the $index(t_j)$ converts the token t_j to its index in the vocabulary of HistBERT model.

2) SENTENCE POSITION EMBEDDING

Sentence position embedding captures the position of each sentence within the input document d. This type of embedding allows the model to represent the document's structural features. Specifically, given the input document $d = \{S_1, S_2, \ldots, S_n\}$, we encode each sentence S_i according to its position. As defined in Equation 2, each token t_j in the *i-th* sentence S_i is encoded as a $d_{Sentpos}$-dimensional vector denoted as $EMB_{Sentpos}(t_j)$, where each element is set to the position of the sentence S_i.

$$EMB_{Sentpos}(t_j) = [pos_{S_i}, pos_{S_i}, \ldots, pos_{S_i}] \tag{2}$$

where pos_{S_i} is the position of the sentence S_i in the document d. For instance, each token t_j in the first sentence will be encoded using the vector $EMB_{Sentpos}(t_j) = [1, 1, \ldots, 1]$, where the dimension $d_{Sentpos}$ is the same as the token embedding dimension.

3) TOKEN POSITION EMBEDDING

Token position embedding encodes the position of each word within its respective sentence. By incorporating token position embedding, a text summarization model can better understand the relationships between words in the input sentences. This type of embedding is particularly useful for capturing syntactic information and understanding the grammatical structure of the input text. Moreover, many studies have shown that it is crucial for text summarization tasks to capture the relative position of each token [44], [56]. As done in previous work for biomedical text summarization [44], we determine the position embedding of each token t_j in the sentence S_i using a fixed lookup table that maps the position of each token into a d_{Tokpos}-dimensional vector, formally defined in the following Equation:

$$\begin{aligned} EMB_{Tokpos}(t_j) = [&EMB_{Tokpos}(p_j, 1), \ldots, \\ &EMB_{Tokpos}(p_j, d_{Tokpos})] \end{aligned} \tag{3}$$

where $EMB_{Tokpos}(t_j)$ denotes the position embedding of the token t_j, $EMB_{Tokpos}(p_j, m)$ is the value of the $m-th$ dimension in vector $EMB_{Tokpos}(t_j)$ computed using the Equation 4, and p_j is the position of of the token t_j in the document d.

$$\begin{cases} EMB_{Tokpos}(p_j, 2m) = sin(\frac{p_j}{10000^{2m/d_{Tokpos}}}) \\ \\ EMB_{Tokpos}(p_j, 2m+1) = cos(\frac{p_j}{10000^{2m/d_{Tokpos}}}) \end{cases} \tag{4}$$

The model uses trigonometric functions to map a token's position to a position embedding. ***Sine*** and ***cosine*** functions are used to compute even and odd dimensions of the embedding. The absolute position information is provided by the token's position, while the periodicity of the trigonometric functions allows different token embeddings to express each other in any dimension. This helps the model learn the relative position relationship between tokens effectively.

4) FUSION EMBEDDING

The fusion embedding vector of each token t_j is the sum of the three embedding vectors, including sentence token embedding, sentence position embedding, and token position embedding, formally calculated in the following Equation:

$$EMB_{final}(t_j) = EMB_{token} + EMB_{Sentpos} + EMB_{Tokpos} \tag{5}$$

During the fine-tuning stage, the attention layer of the HistBERT encoder is fed with the $EMB_{final}(t_j)$ embedding, which is then fine-tuned to improve its ability to represent the contextual information more accurately.

Using the fine-tuned HistBERTSum encoder, each sentence S_i in the input document d is then represented by a contextual embedding vector T_i, where $1 \leq i \leq n$ and T_i corresponds to the vector of the *i-th* [CLS] symbol from the top layer. It is worth noticing that the sinusoid positional embeddings function [1] has been added to indicate the position of each sentence.

B. ABSTRACTIVE SUMMARIZATION PROCESS

Given a historical document d, we use an encoder-decoder model based on the transformer architecture [1] to generate an abstractive summary *Sum* for the input document d. The encoder is the pre-trained HistBERTSum model, and the decoder is a transformer with 6 randomly initialized layers. It is worth mentioning that the HistBERTSum encoder is already pre-trained while the decoder needs to be trained from scratch. Hence, a mismatch between the encoder and decoder may lead to instability during fine-tuning, such as overfitting the encoder and underfitting the decoder or vice versa. To overcome this issue, we use two different Adam optimizer values with $\beta_1 = 0.9$ and $\beta_2 = 0.999$ for the encoder and the decoder, respectively. Additionally, we use different warmup steps and learning rates for each optimizer during training, formally defined in the following Equation:

$$\begin{aligned} lr_E &= \tilde{lr}_E.min(step^{-0.5}, step.warmup_E^{-1.5}) \\ lr_D &= \tilde{lr}_D.min(step^{-0.5}, step.warmup_D^{-1.5}) \end{aligned} \tag{6}$$

We set the encoder's warm-up settings and learning rates to $warmup{-}E = 20,000$ and $\tilde{lr}_E = 2e^{-3}$, respectively, and the decoder's warm-up settings and learning rates to $warmup{-}D = 10,000$ and $\tilde{lr}_D = 0.1$, respectively. With these settings, the pre-trained HistBERTSum encoder is fine-tuned with a lower learning rate, while the transformer decoder is learned from scratch until it becomes stable. Furthermore, we use the trigram blocking technique [57] during the prediction process to avoid certain trigrams from appearing in the summary. The idea behind this is to prevent the summary from being too similar to the original text and ensure that the summary is coherent and semantically meaningful. In addition, this promotes diverse and informative summaries by encouraging the model to use alternative wordings and phrasings. Noticing that we focus on building a minimum-requirements system.

IV. EXPERIMENTAL RESULTS

In this section, we first cover the dataset collection process, the evaluation measure, and the experimental setup. Then, we present and analyze the obtained results, including an ablation analysis of our proposed approach.

A. DATASET COLLECTION

As previously stated, a major challenge for natural language processing in the historical domain is the lack of annotated datasets required for training any supervised machine learning model. Therefore, creating a high-quality summarization dataset is essential to achieve accurate and reliable results. To our knowledge, no dataset currently exists for historical text summarization. To address this gap, we constructed a dataset specifically for historical English documents. Our dataset was obtained from the *Centre Virtuel de la Connaissance sur l'Europe (CVCE) group*,[10] which provides an innovative perspective on Europe's development. We collected 7800 English documents on various topics, such as news, politics, and interviews, written by CVCE experts. We downloaded the documents as PDFs and extracted the required contents, such as the document ID, title, summary, page content, and number of pages. After analyzing the availability of content and summaries across the 7800 documents, we found that only 5761 documents contained both and chose them for further processing. Specifically, the dataset contains 5761 historical documents accompanied by their highlights, answering the question ***"What is this article about?"***. The highlights are designed to capture the main idea of the documents. While these summaries are concise, this approach aligns with established practices in abstractive summarization datasets, such as the *CNN/DailyMail* dataset.[11] Indeed, Nallapati et al. [8] justify the use of short summaries as an effective way to represent document content. Although the highlights are brief, they aim to provide a general, topic-focused summary of the document rather than a detailed narrative. This design is intentional and aligns with the objectives of abstractive summarization tasks, which prioritize capturing the most critical information over providing exhaustive details.

Table 1 summarizes the number of historical documents based on the availability of the content and summary.

After further analysis of the 5761 documents, we found that there was no content balance among them. Some documents had only a few sentences, while others had more than 200 sentences. Similarly, some documents were only 10-15 pages long, while others were much longer. To address this issue, we applied additional filters based on factors such as the number of words in the abstract and content, the minimum sentence length, and the number of valid sentences. We considered a document to be valid if it contained between 15 to 150 sentences in the content, as shown in Figure 3. Additionally, we selected documents with a small number of pages ($N \leq 10$), as shown in Figure 4. After applying these filters, we identified 3907 documents that met our criteria for quality and content balance. The resulting documents had an average of 46 sentences, 1575 words, and four pages. To prepare the dataset for training, we divided these documents into three sets: a training set containing 3163 documents, a test set containing 372 documents, and a validation set containing 372 documents. Furthermore, we applied topic modeling, specifically BERTopic—a state-of-the-art technique [58]—to our corpus to explore and analyze the topic distributions within the collection. This process identified a total of 24 distinct topics, offering a comprehensive overview of the underlying themes in the dataset. These topics provide valuable insights into the structure and key areas of focus within the collection. Figure 2 illustrates the distribution of the top 10 topics in the corpus, based on the number of documents associated with each topic. This visualization highlights the most common topics, providing an intuitive view of their prevalence in the dataset.

TABLE 1. Number of historical documents based on the summary and content availability.

Summary	Content	No. of Documents
Available	Available	5761
Not Available	Available	1931
Not Available	Not Available	94
Available	Not Available	32

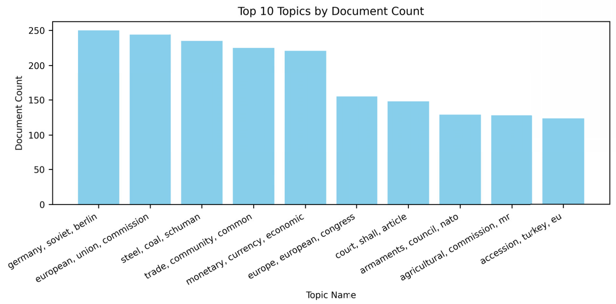

FIGURE 2. Top 10 topics based on document count.

B. EVALUATION MEASURES

To evaluate the effectiveness of the proposed method, we used ROUGE (Recall-Oriented Understudy for Gisting Evaluation) [41]. Specifically, we have used ROUGE-N (ROUGE-1 and ROUGE-2) and ROUGE-L. ROUGE-N

[10] https://www.cvce.eu/

[11] https://huggingface.co/datasets/abisee/cnn_dailymail

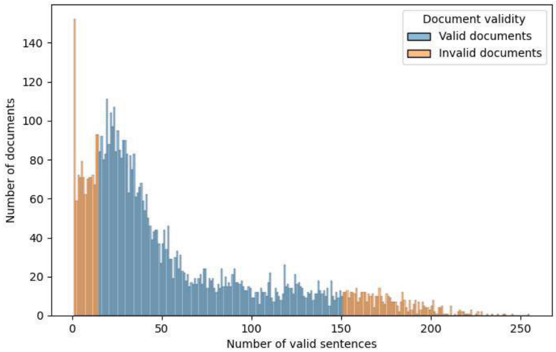

FIGURE 3. Distribution of total documents based on number of valid sentences.

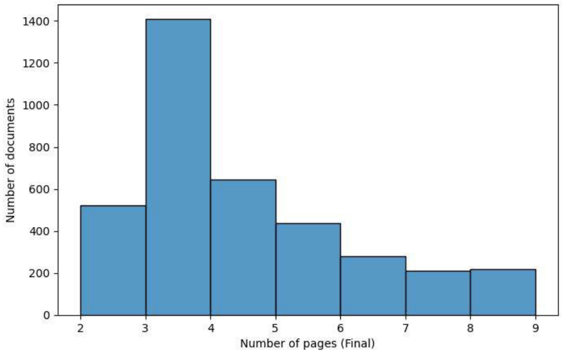

FIGURE 4. Distributions of total documents based on number of pages (N < = 10).

determines the similarity between the systems summaries and a set of gold summaries based on the n-gram overlap, whereas ROUGE-L evaluates the fluency of the summary. It is based on the Longest Common Subsequence (LCS) that takes into account sentence-level structure similarity. We have reported the obtained F1 performance of ROUGE-1 (R-1), ROUGE-2 (R-2), and ROUGE-L (R-L) using the official ROUGE toolkit (version 1.5.5) with standard options settings used for assessing abstractive single document summarization systems. It is worth noting that the ROUGE method focuses specifically on the informativeness of the produced summary. In fact, a recent study [59] has demonstrated that no other automatic metric consistently outperforms the ROUGE method in evaluating text summarization systems.

C. EXPERIMENTAL SETUP

The HistBERTSum-Abs method is implemented using PyTorch and is based on the `'bert-base-uncased'`[12] version of the BERT model. Input documents were tokenized using BERT's subword tokenizer. The model was trained on the high-performance Iris cluster[13] at the University of Luxembourg, which features 96 Nvidia V100 GPU-AI accelerators with Skylake or Broadwell processors. Specifically, we utilized 4 GPUs with ten cores and one node. The architecture consists of six Transformer layers, each containing 768 hidden units. Dropout with a probability of 0.1 was applied before all linear layers. Additionally, the trigram blocking method [57] was employed to prevent the generation of repeated trigrams in the produced summaries. The HistBERTSum-Abs model was trained for 250,000 steps with gradient accumulation every five steps, saving check-points every 5,000 steps, and evaluating on the validation set. During decoding, we used a beam search with a beam size of 5 and tuned the α parameter for the length penalty [60] between 0.6 and 1 on the validation set. Decoding continued until the end-of-sequence token was generated. It is worth noting that the use of a subword tokenizer in the proposed method ensures that out-of-vocabulary words rarely occur in the output. Furthermore, the incorporation of trigram blocking enhances the diversity of the generated summaries and effectively reduces repetition.

D. COMPARISON WITH STATE-OF-THE-ART METHODS

We evaluated the effectiveness of the proposed method, HistBERTSum-Abs, by comparing its performance with recent state-of-the-art (SOTA) methods for abstractive single-document summarization on our historical dataset. The ROUGE F1 scores (R-1, R-2, and R-L) achieved by HistBERTSum-Abs and the SOTA systems are presented in Table 2. The first block of the table shows the ROUGE F1 scores of two extractive baselines: ORACLE and LEAD-3. ORACLE represents an extractive upper bound using a sentence selection technique that maximizes the ROUGE-2 score for the target summaries, while LEAD-3 generates summaries by selecting the first three sentences of each document. The second block of the table summarizes the F1 scores of several abstractive SOTA systems evaluated on our dataset, including abstractive models specifically designed for the legal domain. The third block of the table presents the results obtained by the three selected LLMs under zero-shot settings. Additionally, examples of the generated summaries for all these systems, including our method, are provided in the Appendix VI. These systems are briefly described below:

- **BERTSUM [14]** is a single document summarization model for extractive and abstractive approaches, which involves using a new BERT-based document-level encoder to capture document semantics and create sentence-level representations.
- **BART [15]** is a denoising auto-encoder for pretraining sequence-to-sequence models, which involves corrupting input texts with noise and training the model to reconstruct them.
- **DistBART**[14] is a modified version of BART [15] model, which is trained on the CNN/DailyMail [8] and XSum [61] datasets for abstractive single-document summarization using the knowledge distillation approach.

[12] https://github.com/huggingface/transformers

[13] https://hpc-docs.uni.lu/systems/iris/

[14] https://huggingface.co/sshleifer/distilbart-cnn-12-6/tree/main

- **T5 [16]** is a transformer-based language model developed by Google that can be fine-tuned for various natural language processing tasks, including text classification, summarization, and question-answering.
- **PEGASUS [17]** is a state-of-the-art transformer-based language model developed by Google, specifically designed for abstractive text summarization tasks. It has achieved state-of-the-art performance on several benchmarks.
- **Legal-T5** is a model based on T5-small [16], specifically fine-tuned for the task of legal abstractive summarization. It leverages the JRC-ACQUIS dataset,[15] which contains approximately 22,000 legal documents.
- **Legal-LED** is a model fine-tuned on base-LED [51], specifically tailored for the legal domain and trained for long-document abstractive summarization tasks. It utilizes the SEC Litigation Releases dataset,[16] which contains over 2,700 litigation releases and complaints.
- **Legal-Pegasus** is a fine-tuned version of Pegasus [17], specifically designed for abstractive summarization tasks in the legal domain. Similar to Legal-LED, this model is trained on the SEC Litigation Releases dataset,[17] which contains over 2,700 litigation releases and complaints.

TABLE 2. ROUGE-1 (R-1), ROUGE-2 (R-2), ROUGE-L (R-L) F1 score results of the proposed method -HistBERTSum-Abs, the SOTA abstractive methods and LLM's using our historical dataset.

Model	R-1	R-2	R-L
ORACLE	33.60	15.81	24.94
LEAD-3	30.13	11.84	22.29
Abstractive Models			
BERTSUM	40.83	20.21	37.15
T5	42.53	21.04	37.46
BART	43.24	21.14	37.84
DistBART	43.32	21.25	37.65
PEGASUS	44.02	21.35	38.19
Legal-T5	23.36	07.04	20.61
Legal-LED	46.12	25.26	36.86
Legal-Pegasus	44.04	22.97	35.25
LLMs with Zero Shot			
ChatGPT-4o mini	32.18	09.70	20.89
Llama 3.1 8B Instruct	33.02	11.05	22.22
Mistral-7b-instruct-v0.3	32.21	11.43	21.49
Our Method			
HistBERTSum-Abs	**47.62**	**25.54**	**38.47**

As shown in Table 2, our HistBERTSum-Abs method outperformed the extractive baselines, including ORACLE and LEAD-3, across all evaluation measures. Specifically, the ORACLE system, which selects the best sentences based on ROUGE similarities, and the LEAD-3 baseline, which simply extracts the first three sentences from the input document, scored significantly lower in all ROUGE metrics (R-1, R-2, and R-L). This demonstrates the limitations of extractive approaches in summarizing historical texts, where nuanced contextual understanding is required. Furthermore, our method also outperformed all the other state-of-the-art (SOTA) abstractive systems fine-tuned on our historical dataset. Notably, HistBERTSum-Abs achieved higher scores across all metrics, demonstrating the effectiveness of domain-specific fine-tuning. For example, when compared to PEGASUS, a prominent SOTA system for abstractive single-document summarization, our method achieved improvements of 3.6%, 4.19%, and 0.26% in R-1, R-2, and R-L metrics, respectively. This underscores the ability of HistBERTSum-Abs to generate concise, coherent, and contextually accurate summaries in the historical domain.

We further fine-tuned three recent abstractive summarization models tailored for the legal domain—Legal-T5, Legal-LED, and Legal-Pegasus—on our historical dataset to evaluate their adaptability. As detailed in Table 2, Legal-LED achieved the highest scores among these legal domain models, likely due to its ability to handle long documents effectively. However, it still fell short of the performance of HistBERTSum-Abs, highlighting the challenges of adapting models designed for other domains to the historical domain, even with fine-tuning.

In addition to these experiments, we evaluated the zero-shot performance of large language models, including ChatGPT-4o Mini, Mistral-7B-Instruct, and Llama 3.1 8B Instruct. As shown in the third block of Table 2, these LLMs produced suboptimal ROUGE scores compared to our fine-tuned approach. However, despite their lower ROUGE scores, the zero-shot LLMs were able to generate coherent and semantically rich summaries. This observation highlights a potential limitation of ROUGE as an evaluation metric, as it primarily measures lexical overlap rather than the quality of semantic coherence or contextual understanding. The results, as demonstrated in the Appendix VI, show that LLM-generated summaries often captured key ideas effectively but did not align word-for-word with the reference summaries. This discrepancy underscores the need for further research into evaluation methods for summarization tasks, particularly metrics that can better account for semantic fidelity and contextual accuracy.

Furthermore, as illustrated in the Appendix VI, the summaries generated by HistBERTSum-Abs are concise and contextually accurate. These results validate the hypothesis that utilizing the HistBERT model, fine-tuned on a historical domain-specific corpus, enhances the performance of text summarization in the historical domain. Additionally, these findings underscore the critical importance of fine-tuning pre-trained language models on domain-specific datasets to improve their effectiveness. This argument is further supported by recent research [62], [63], which demonstrates through experimental evidence that LLMs tend to perform poorly on domain-specific data in zero-shot and few-shot settings without proper adaptation.

E. MODEL ANALYSIS

In addition to comparing our method to several state-of-the-art systems, we conducted self-validation by analyzing

[15] https://wt-public.emm4u.eu/Acquis/index_2.2.html

[16] https://www.sec.gov/enforcement-litigation/litigation-releases

[17] https://www.sec.gov/enforcement-litigation/litigation-releases

different components of our approach, as described in the subsequent subsections.

1) EFFECT OF EMBEDDING

As already mentioned, the sentence representation is constructed using the summation of the three embedding layers: sentence token embedding, sentence position embedding, and token position embedding. We tested the effectiveness of sentence position and token position embeddings by conducting three scenarios on our historical dataset. In the first scenario, we removed the token position layer, and the input representation was the sum of the sentence token and sentence position embeddings. In the second scenario, we removed the sentence position embedding layer. In the third scenario, we used all three embedding layers. Table 3 summarizes the ROUGE results obtained from these scenarios. From Table 3, it seems clear that there is a significant decrease in ROUGE scores when any of the embedding layers is removed. Therefore, this analysis confirms that the use of all three embeddings is crucial in capturing a better sentence representation, which is necessary for generating a high-quality summary. The comparison results further demonstrate that the three used embeddings are complementary to each other.

TABLE 3. Effect of removing token position and sentence position embeddings layers with regards to ROUGE F1 scores.

Effect of Embedding Layers	R-1	R-2	R-L
Removing Token Position Layer	40.02 ↓	20.04 ↓	32.75 ↓
Removing Sentence Position Layer	45.86 ↓	24.05 ↓	36.19 ↓
Combining All the three Layers	**47.62**	**25.54**	**38.47**

2) EFFECT OF FINE-TUNING THE HISTBERT ENCODER

In this analysis, we aim to evaluate the effectiveness of fine-tuning the HistBERT encoder for our HistBERTSum-Abs summarization system. We conducted two experiments: i) we fine-tuned only the abstractive summarization layers, and ii) we fine-tuned both the encoder and the summarization layers. The results presented in Table 4 showed a decrease in ROUGE scores when we turned off fine-tuning for the encoder. The obtained results can be explained by the fact that fine-tuning the HistBERT encoder on the summarization task further adapts it to the specific task at hand. The HistBERT encoder is the backbone of the HistBERTSum-Abs system, and it plays a crucial role in encoding the input text into a dense representation that better captures language patterns and semantic information. Fine-tuning the encoder on our historical dataset improves its ability to represent input text for summarization, resulting in the generation of high-quality summaries. Furthermore, jointly fine-tuning the encoder and the summarization layers allows them to learn to work together effectively. By updating the encoder parameters during training, the summarization layers learn to use the encoder's output better, and the encoder learns to generate better representations that are more suitable for the summarization task.

TABLE 4. Effect of fine-tuning the parameters of HistBERT encoder with regards to ROUGE F1 scores.

Effect of Finetuning HistBERT	R-1	R-2	R-L
Without HistBERT fine-tuning	45.62 ↓	23.52 ↓	37.01 ↓
With HistBERT Fine-tuning	**47.62**	**25.54**	**38.47**

3) EFFECT OF DOMAIN KNOWLEDGE

This analysis aims to evaluate the impact of domain-specific pre-training on the quality of the generated summaries. While the original BERT model is pre-trained on a large general corpus, it may lack domain-specific knowledge. In contrast, HistBERT is further pre-trained on a large corpus of historical documents, COHA.[18] To assess the effectiveness of domain-specific pre-training, we replaced the HistBERT encoder with vanilla BERT and compared the obtained ROUGE scores. The results presented in Table 5 show a significant decrease in ROUGE scores when the vanilla BERT is used in terms of all of the evaluation measures (R-1, R-2, and R-L). We also evaluated our method on test documents from the CNN/Daily Mail dataset and observed significantly poor ROUGE scores. This underscores the critical role of domain knowledge in our approach, which has been fine-tuned on the COHA corpus's historical linguistic characteristics. While our method excels in generating summaries enriched with historical context—including dates, names of key individuals, and other historically relevant details—it diverges from the CNN/Daily Mail gold-standard summaries written in plain contemporary text. This finding confirms that further pre-training a model on domain-specific corpora improves its ability to construct better representations, leading to the generation of more informative summaries.

TABLE 5. Effect of domain knowledge with regards to ROUGE F1 scores.

Effect of Domain Knowledge	R-1	R-2	R-L
Vanilla-BERTSum	↓ 40.83	↓ 20.21	↓ 37.15
HistBERTSum-Abs (CNN)	**↓14.16**	**↓ 01.70**	**↓ 10.57**
HistBERTSum-Abs	**47.62**	**25.54**	**38.47**

4) NUMBER OF NOVEL N-GRAMS

In this analysis, our objective is to compare the number of new n-grams generated in the final summaries of our HistBERTSum-Abs model with the BERTSUMAbs [14] system and the reference summaries for different N-grams (N = 1 to 4). As shown in Figure 5, we found that the proportion of new n-grams generated by the BERTSUMAbs system was smaller than our model and the reference summary. Our proposed model, HistBERTSum-Abs, produced a better proportion of novel n-grams, and the gap between the reference summary was also reduced. These results

[18] https://www.english-corpora.org/coha/

demonstrate that the HistBERTSum-Abs model had a better understanding of the context of the document and was able to generate summaries with new words not present in the source document.

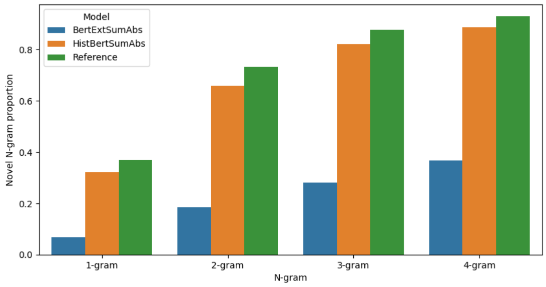

FIGURE 5. Proportion of novel n-grams.

V. CONCLUSION

In this paper, we proposed a novel method called HistBERTSum-Abs for abstractive historical single-document summarization, which is based on the HistBERT encoder [18] - a domain-aware language model. We introduced an effective text representation method that consists of three embedding layers, including the sentence token, sentence position, and token position embeddings. Additionally, to address the lack of historical text summarization datasets, we collected documents from the *Centre Virtuel de la Connaissance sur l'Europe (CVCE) group* and created our own dataset for historical text summarization. The comparison of several recent abstractive text summarization systems, fine-tuned on our historical dataset, as well as SOTA LLMs in zero-shot experiments, demonstrates the effectiveness of the proposed method. Specifically, the use of the HistBERT encoder has shown to be effective for the historical text summarization task. As far as we know, this is the first attempt at historical text summarization in English, which could be a good starting point for future research in this field.

Supervised text summarization relies on high-quality labeled training data for successful model training, which can be cumbersome to acquire, especially for specific domains that require expert annotation. Recent deep learning research has focused on self-supervised learning methods to reduce the need for supervision. Despite the success of fine-tuning deep pre-trained language models in many NLP applications, their generalization performance may suffer from domain shifts. To deal with this issue, we plan to investigate the potential of *unsupervised domain adaptation* and *contrastive learning* approaches for historical text summarization tasks.

VI. EXAMPLES OF GENERATED SUMMARIES USING OUR METHOD-HISTBERTSUM-ABS AND SOTA METHODS

A. EXAMPLE 1

1) SOURCE TEXT

(16 July 1998) A4-0271/1998 The European Parliament, — having regard to the Amsterdam Treaty, — having regard to the 1993 interinstitutional agreement on Article 189b of the EC Treaty concerning the phase preceding the adoption of a common position by the Council and on arrangements for the proceedings of the Conciliation Committee under Article 189b (1), — having regard to Rule 148 of its Rules of Procedure, — having regard to the report of its Committee on Institutional Affairs and the opinions of the Committee on Research, Technology, Development and Energy, the Committee on Legal Affairs and Citizens' Rights, the Committee on the Environment, Public Health and Consumer Protection, the Committee on Culture, Youth, Education and the Media and the Committee on Civil Liberties and Internal Affairs (A4-0271/1998), A. Whereas the Amsterdam Treaty not only extended but also reformed the application of the co-decision procedure, in that it has been simplified and the European Parliament's position is now equal to that of the Council, as is appropriate in a two-chamber legislative system, B. Whereas implementing measures will have to be taken if the full potential of the new co-decision procedures is to be realized, in particular by changing the nature of interinstitutional relations in the first reading and by further improvements in the second reading and conciliation phases, C. Whereas such implementing measures can partly be carried out by reform of Parliament's own procedures, Parliament's Rules of Procedure should therefore be amended, D. However, these internal reforms need to be complemented by procedural changes reached in agreement with the Commission and the Council, E. Whereas such changes will require the 1993 interinstitutional agreement on co-decision to be updated and extended in scope, First reading phase 1. Points out that the Amsterdam Treaty provides for a major change in the functioning of the co-decision procedure by permitting it to be concluded at the first reading stage, thereby accelerating, rationalizing, and simplifying the EU legislative process; 2. Believes that if new rules of procedure are to be effectively implemented, a structured dialogue between Parliament, the Council, and the Commission is also necessary.

2) GOLD SUMMARY

European Parliament resolution of 16 July 1998 on the new co-decision procedure after Amsterdam.

3) MODEL SUMMARY

1) **HistBERTSum-Abs (OURS)**: European Parliament resolution of 16 July 1998 incorporating parliament's opinion on the Amsterdam treaty. The European Parliament recognizes the major change in the functioning of the co-decision procedure since the 1993 interinstitutional agreement on the revision of the EC treaty.
2) **BART**: On 16 July 1998, a4 - 0271/1998, the European parliament, having regard to the Amsterdam Treaty, — having regard to the 1993 interinstitutional agreement on article 189b of the EC treaty concerning the phase preceding the adoption of a common position by the

council and on arrangements for the proceedings of the conciliation committee under article 188b (1)

3) **DistBART**: On 16 July 1998, the European Parliament adopts a resolution in which it considers the implementation of the Amsterdam treaty on the procedure of the EC treaty concerning the phase preceding the adoption of a common position by the council and on arrangements for the proceedings of the conciliation committee under article 189b (1) of the treaty on ec treaty<q>the EU parliament's rules of procedure (a4 - 0271/1998) should be amended in order to improve the procedures of its parliamentary committees, since the protocol on the role of the national parliaments in the EU.
4) **PEGASUS**: European parliament resolution of 16 July 1998 on the practical application of the new co-decision procedure arising from the agreement on 'Amsterdam treaty ', signed on 7 December 1993
5) **T5**: On 16 July 1998, the European parliament, having regard to the Amsterdam treaty, considers that the new co-decision procedure should be both updated and extended in scope
6) **Legal_T5**: On 4 May 1969, the council of the European union (weu) circula of the European union (weu) circula of the council of the European union (weu), the European parliament on the council of the European union
7) **Legal_pegasus**: On 16 July 1998, the European parliament adopts a resolution (a4 - 0271/1998) in which it emphasises the need for a structured dialogue between the council, the European Commission and the national parliaments in order to implement the new rules of procedure
8) **Legal_LED**: On 16 July 1998, the European parliament adopts a resolution in which it calls for a major change in the functioning of the co-decision procedure by enabling it to be concluded at first reading stage, thereby accelerating, rationalizing and simplifying the EU legislative process
9) **ChatGPT-4o Mini**: On July 16, 1998, the European Parliament discussed the implications of the Amsterdam Treaty on the codecision procedure, emphasizing the need for procedural reforms to enhance interinstitutional relations and legislative efficiency. The Parliament called for updated interinstitutional agreements and specific measures to improve transparency, legal quality, and collaboration between the Parliament, Council, and Commission during the legislative process.
10) **Mistral-7B-Instruct-v0.3**:The document discusses the need for implementing measures to fully utilize the potential of the new codecision procedures in the European Union, as reformed by the Amsterdam Treaty. It suggests changes in parliament's own procedures, as well as procedural changes in agreement with the commission and the council. The document proposes steps to improve first reading, second reading, and conciliation phases, including structured dialogue between institutions, improved legal quality of texts.
11) **Llama 3.1 8B Instruct**: The European Parliament proposes several measures to improve the legislative process, including the creation of a structured dialogue between parliament, council, and commission, the use of informal trialogues, and the exchange of documents between institutions. The parliament also calls for the revision of the existing interinstitutional agreement on codecision to take into account the changes made in the Amsterdam treaty.

B. EXAMPLE 2

1) SOURCE TEXT

(30 November 2006) The European Parliament, — having regard to the Treaty of Accession of the Republic of Bulgaria and Romania to the European Union(1), signed on 25 April 2005, — having regard to all its previous resolutions and reports since the beginning of the enlargement process, in particular its most recent resolution of 14 June 2006 on the accession of Bulgaria and Romania (2), — having regard to the Presidency conclusions of the Brussels European Council of 15/16 June 2006, — having regard to the conclusions of the General Affairs and External Relations Council on enlargement of 17 October 2006, — having regard to the Commission's monitoring report on Romania of 26 September 2006 (COM(2006)0549) and its previous monitoring reports, — having regard to the exchange of letters between the President of the European Parliament and the President of the Commission on the full involvement of the European Parliament in any consideration of activation of one of the safeguard clauses in the Treaty of Accession, — having regard to Rule 45 of its Rules of Procedure, — having regard to the report of the Committee on Foreign Affairs and the opinions of the Committee on Civil Liberties, Justice and Home Affairs and the Committee on Women's Rights and Gender Equality (A6- 0421/2006), A. Whereas Romania's accession to the European Union is a major historical development, accompanied by a profound change in the economic, social, and civil landscape of the country, and this accession will have positive effects on the Romanian population and on the development and cohesion of the European Union, B. Whereas the integration of Romania into the European Union will contribute to the stability and prosperity of south-east Europe, C. Whereas the incorporation of Romania into the European Union will strengthen the political and cultural dimension of the European integration process, D. Whereas the first stage of the fifth enlargement in 2004 proved beneficial for both the old and new Member States, this will undoubtedly also be true of the current enlargement, bringing the fifth enlargement to a conclusion, E. Since the Commission report in May 2006, further significant improvements can be noted, as detailed in the Commission's most recent monitoring report of 26 September 2006, F. Whereas Romania is

continuing its efforts to fulfill the conditions set in the Treaty of Accession to become a member of the Union on 1 January 2007, with Bulgaria, Parliament has consistently expressed its desire to see these two countries join simultaneously, 1. Commends the Commission on the conscientious and rigorous nature of its work on monitoring the reforms undertaken by Romania; 2. Congratulates Romania and welcomes its accession on 1 January 2007, looks forward to the arrival in due 2 / 5 05/09/2012 institutions, and recognizes the excellent contribution made by Romania's Observers in the European Parliament since September 2005; 3. Highlights the fact that the current enlargement of the European Union, like previous ones, embodies the ideas of European unity and solidarity, benefiting all parties and making it possible to promote the values of democracy, equality, pluralism, and non-discrimination; 4. Welcomes the fact that the Commission's report of 26 September 2006 recommended that Bulgaria and Romania should join simultaneously.

2) GOLD SUMMARY

On 30 November 2006, the European Parliament adopted its final resolution on Bulgaria's accession to the European Union.

3) MODEL SUMMARY

1) **HistBERTSum-Abs (OURS)**: On 30 November 2006, the European Parliament adopted a resolution welcoming Bulgaria and Romania's accession to the European Union, effective 1 January 2007. The resolution highlights the importance of their efforts to meet EU standards and emphasizes the benefits of this enlargement for European unity and shared values.
2) **BART**: On 30 November 2006, the European parliament welcomes the accession of Bulgaria and Romania to the EU.
3) **DistBART**: On 30 November 2006, the European parliament adopts a resolution on the accession of the Republic of Bulgaria and Romania to the EU on 1 January 2007.The resolution recognises the progress of the reform process in the country since the opening of accession negotiations and emphasises the importance of continuing to involve the commission in monitoring developments in the monitoring of developments.
4) **PEGASUS**: On 30 November 2006, the European Parliament adopts a resolution in which it welcomes the confirmation by the commission and the council that Bulgaria is sufficiently prepared for accession to the European Union on 1 January 2007
5) **T5**:On 30 November 2006, the European Parliament publishes a monitoring report on the state of preparedness for accession of Bulgaria to the European union.
6) **Legal_T5**: On 25 October 1949, the German daily newspaper's accession to the European union.
7) **Legal_pegasus**: On 30 November 2006, the European Parliament adopts a resolution in which it welcomes the accession of Bulgaria and Romania to the European union and calls on the Republic of Bulgaria to continue its reform process
8) **Legal_LED**: On 30 November 2006, the European parliament adopts a resolution on the state of preparedness for the accession of Bulgaria and Romania to the European union
9) **ChatGPT-4o Mini**: On November 30, 2006, the European Parliament welcomed Bulgaria's confirmation of readiness to join the EU on January 1, 2007, following a comprehensive evaluation of its accession process and reforms. The Parliament emphasized the need for continued reforms and monitoring to ensure Bulgaria's successful integration into the EU, while also acknowledging its significant economic progress and contributions to regional stability.
10) **Mistral-7B-Instruct-v0.3**: The European Parliament congratulates Bulgaria on its accession to the European Union on 1 January 2007, recognizing its significant achievements during the transformation process. The Parliament highlights areas of concern, such as the need for continued reforms, transparency in governance, and addressing shortcomings in justice and home affairs, justice reform, and the protection of minorities. The Parliament also calls for the proper management and control of EU funds, and for the proper functioning of EU
11) **Llama 3.1 8B Instruct**: The European Parliament congratulates Bulgaria on its accession to the European Union on January 1, 2007, and welcomes the country's progress in meeting the necessary reforms. The Parliament notes areas requiring further progress, including the fight against organized crime, corruption, and money laundering, and calls on the Bulgarian authorities to act with urgency and thoroughness.

ACKNOWLEDGMENT

The authors would like to acknowledge Dr. Lars Wieneke and his team from the Digital Infrastructure Department at Luxembourg Centre for Contemporary and Digital History, Luxembourg University, for providing the dataset used in this study.

REFERENCES

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, Jun. 2017, pp. 5998–6008.

[2] M. M. Terras, "The rise of digitization," in *Digitisation Perspectives. Educational Futures Rethinking Theory and Practice*, vol. 46, R. Rikowski, Ed., SensePublishers, 2011, doi: 10.1007/978-94-6091-299-3_1.

[3] S. A. South, *Method and Theory in Historical Archeology* (Institute for Research on Poverty Monograph Series), 1st ed. Academic Press, Jan. 1977.

[4] H. P. Luhn, "A statistical approach to mechanized encoding and searching of literary information," *IBM J. Res. Develop.*, vol. 1, no. 4, pp. 309–317, Oct. 1957.

[5] V. Gupta and G. S. Lehal, "A survey of text summarization extractive techniques," *J. Emerg. Technol. Web Intell.*, vol. 2, no. 3, pp. 258–268, Aug. 2010.

[6] R. Ferreira, L. de Souza Cabral, R. D. Lins, G. Pereira e Silva, F. Freitas, G. D. C. Cavalcanti, R. Lima, S. J. Simske, and L. Favaro, "Assessing sentence scoring techniques for extractive text summarization," *Exp. Syst. Appl.*, vol. 40, no. 14, pp. 5755–5764, Oct. 2013.

[7] S. Lamsiyah, A. E. Mahdaouy, S. E. A. Ouatik, and B. Espinasse, "Unsupervised extractive multi-document summarization method based on transfer learning from BERT multi-task fine-tuning," *J. Inf. Sci.*, vol. 49, no. 1, pp. 164–182, Feb. 2023.

[8] R. Nallapati, B. Zhou, C. dos Santos, C. Gulcehre, and B. Xiang, "Abstractive text summarization using sequence-to-sequence RNNs and beyond," in *Proc. 20th SIGNLL Conf. Comput. Natural Lang. Learn.*, 2016, pp. 280–290.

[9] H. Lin and V. Ng, "Abstractive summarization: A survey of the state of the art," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, Jul. 2019, pp. 9815–9822.

[10] S. Liu, J. Cao, R. Yang, and Z. Wen, "Key phrase aware transformer for abstractive summarization," *Inf. Process. Manage.*, vol. 59, no. 3, May 2022, Art. no. 102913.

[11] Y. Gong and X. Liu, "Generic text summarization using relevance measure and latent semantic analysis," in *Proc. 24th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Sep. 2001, pp. 19–25.

[12] S. Fisher and B. Roark, "Query-focused summarization by supervised sentence ranking and skewed word distributions," in *Proc. Document Understand. Workshop (DUC)*, New York, NY, USA, Jun. 2006, p. 8.

[13] 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. Conf. North American Chapter Assoc. Comput. Linguistics, Human Language Technol.*, vol. 2, Jun. 2018, pp. 615–621.

[14] Y. Liu and M. Lapata, "Text summarization with pretrained encoders," in *Proc. Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process. (EMNLP-IJCNLP)*, 2019, pp. 3730–3740.

[15] 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 Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 7871–7880.

[16] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *J. Mach. Learn. Res.*, vol. 21, no. 1, pp. 5485–5551, 2020.

[17] J. Zhang, Y. Zhao, M. Saleh, and P. J. Liu, "PEGASUS: Pre-training with extracted gap-sentences for abstractive summarization," in *Proc. Int. Conf. Mach. Learn.*, Jan. 2019, pp. 11328–11339.

[18] W. Qiu and Y. Xu, "HistBERT: A pre-trained language model for diachronic lexical semantic analysis," 2022, *arXiv:2202.03612*.

[19] Z. Nasar, S. W. Jaffry, and M. K. Malik, "Textual keyword extraction and summarization: State-of-the-art," *Inf. Process. Manage.*, vol. 56, no. 6, Nov. 2019, Art. no. 102088.

[20] W. S. El-Kassas, C. R. Salama, A. A. Rafea, and H. K. Mohamed, "Automatic text summarization: A comprehensive survey," *Exp. Syst. Appl.*, vol. 165, Mar. 2021, Art. no. 113679.

[21] M. Piotrowski, "Natural language processing for historical texts," *Synth. Lectures Human Lang. Technol.*, vol. 5, no. 2, pp. 1–157, Sep. 2012.

[22] M. Ehrmann, A. Hamdi, E. L. Pontes, M. Romanello, and A. Doucet, "Named entity recognition and classification on historical documents: A survey," 2021, *arXiv:2109.11406*.

[23] T. Hills and A. Miani. *A Short Primer on Historical Natural Language Processing*. [Online]. Available: https://api.semanticscholar.org/CorpusID:273174599

[24] T. Zhang, F. Ladhak, E. Durmus, P. Liang, K. McKeown, and T. B. Hashimoto, "Benchmarking large language models for news summarization," *Trans. Assoc. Comput. Linguistics*, vol. 12, pp. 39–57, Jan. 2024.

[25] H. Jin, Y. Zhang, D. Meng, J. Wang, and J. Tan, "A comprehensive survey on process-oriented automatic text summarization with exploration of LLM-based methods," 2024, *arXiv:2403.02901*.

[26] J. Holt and A. Perren, *Media Industries: History, Theory, and Method*. Hoboken, NJ, USA: Wiley, 2011.

[27] S. Oepen, K. Hagen, and J. B. Johannessen, Eds., *Proc. 19th Nordic Conf. Comput. Linguistics (NODALIDA)*. Oslo, Norway: Linköping Univ. Electronic Press, Sweden, May 2013. [Online]. Available: https://aclanthology.org/W13-5600/

[28] G. Bouma and Y. Adesam, Eds., *Proc. NoDaLiDa Workshop Process. Historical Lang.* Gothenburg, Sweden: Linköping Univ. Electronic Press, May 2017.

[29] T. Rama, "Studies in computational historical linguistics: Models and analyses," Ph.D. thesis, 2015. [Online]. Available: https://gupea.ub.gu.se/handle/2077/40571?locale-attribute=sv

[30] E. Pettersson, J. Lindström, B. Jacobsson, and R. Fiebranz, "Histsearch-implementation and evaluation of a web-based tool for automatic information extraction from historical text," in *Proc. HistoInformatics@ DH*, 2016, pp. 25–36.

[31] M. Bollmann, A. Søgaard, and J. Bingel, "Multi-task learning for historical text normalization: Size matters," in *Proc. Workshop Deep Learn. Approaches Low-Resource NLP*, 2018, pp. 19–24.

[32] E. T. K. Sang, M. Bollmann, R. Boschker, F. Casacuberta, F. Dietz, S. Dipper, M. Domingo, R. V. D. Goot, M. V. Koppen, N. Ljubešić, R. Östling, F. Petran, E. Pettersson, Y. Scherrer, M. Schraagen, L. Sevens, J. Tiedemann, T. Vanallemeersch, and K. Zervanou, "The CLIN27 shared task: Translating historical text to contemporary language for improving automatic linguistic annotation," *Comput. Linguistics Netherlands J.*, vol. 7, pp. 53–64, Dec. 2017.

[33] Y. Yang and J. Eisenstein, "Part-of-speech tagging for historical English," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Human Lang. Technol.*, 2016, pp. 1318–1328.

[34] B. Blouin, B. Favre, J. Auguste, and C. Henriot, "Transferring modern named entity recognition to the historical domain: How to take the step?" in *Proc. Workshop Natural Language Process. Digit. Humanities*, 2021, pp. 152–162.

[35] W. L. Hamilton, K. Clark, J. Leskovec, and D. Jurafsky, "Inducing domain-specific sentiment lexicons from unlabeled corpora," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2016, pp. 595–605.

[36] R. Sprugnoli and S. Tonelli, "Novel event detection and classification for historical texts," *Comput. Linguistics*, vol. 45, no. 2, pp. 229–265, Jun. 2019.

[37] V. Lai, M. V. Nguyen, H. Kaufman, and T. H. Nguyen, "Event extraction from historical texts: A new dataset for black rebellions," in *Proc. Findings Assoc. Comput. Linguistics*, 2021, pp. 2390–2400.

[38] C. Liebeskind and S. Liebeskind, "Deep learning for period classification of historical Hebrew texts," *J. Data Mining Digit. Humanities*, vol. 2020, 2020, doi: 10.46298/jdmdh.5864.

[39] J. Gung and J. Kalita, "Summarization of historical articles using temporal event clustering," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Human Lang. Technol.*, Jun. 2012, pp. 631–635.

[40] P. P. Ghosh, R. Shahariar, and M. A. H. Khan, "A rule based extractive text summarization technique for Bangla news documents," *Int. J. Modern Educ. Comput. Sci.*, vol. 10, no. 12, pp. 44–53, Dec. 2018.

[41] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Proc. Text Summarization Branches Out*, Jul. 2004, pp. 74–81.

[42] X. Peng, Y. Zheng, C. Lin, and A. Siddharthan, "Summarising historical text in modern languages," in *Proc. 16th Conf. Eur. Chapter Assoc. Comput. Linguistics, Main Volume*, 2021, pp. 3123–3142.

[43] S. Lamsiyah, K. Murugaraj, and C. Schommer, "Historical-domain pre-trained language model for historical extractive text summarization," in *Proc. 8th Int. Conf. Comput. Inf. Sci. Technol.*, Aug. 2023, pp. 1–9.

[44] Y. Du, Q. Li, L. Wang, and Y. He, "Biomedical-domain pre-trained language model for extractive summarization," *Knowl.-Based Syst.*, vol. 199, Jul. 2020, Art. no. 105964.

[45] A. M. Rush, S. Chopra, and J. Weston, "A neural attention model for abstractive sentence summarization," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2015, pp. 379–389.

[46] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North American Chapter Assoc. Comput. Linguistics, Human Lang. Technol.*, vol. 1, Jan. 2018, pp. 4171–4186.

[47] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," in *Proc. Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process. (EMNLP-IJCNLP)*, 2019, pp. 3982–3992.

[48] Z. Liu, W. Lin, Y. Shi, and J. Zhao, "A robustly optimized BERT pre-training approach with post-training," in *Proc. 20th Chin. Nat. Conf. Comput. Linguistics*, 2021, pp. 1218–1227.

[49] A. Shukla, P. Bhattacharya, S. Poddar, R. Mukherjee, K. Ghosh, P. Goyal, and S. Ghosh, "Legal case document summarization: Extractive and abstractive methods and their evaluation," in *Proc. 2nd Conf. Asia–Pacific Chapter Assoc. Comput. Linguistics 12th Int. Joint Conf. Natural Lang. Process.*, 2022, pp. 1048–1064.

[50] D. D. V. Feijo and V. P. Moreira, "Improving abstractive summarization of legal rulings through textual entailment," *Artif. Intell. Law*, vol. 31, no. 1, pp. 91–113, Nov. 2021.
[51] I. Beltagy, M. E. Peters, and A. Cohan, "Longformer: The long-document transformer," 2020, *arXiv:2004.05150*.
[52] A. A. Lins, C. S. Carvalho, F. D. C. J. Bomfim, D. D. C. Bentes, and V. Pinheiro, "CLSJUR.B R—A model for abstractive summarization of legal documents in Portuguese language based on contrastive learning," in *Proc. 16th Int. Conf. Comput. Process.*, P. Gamallo, D. Claro, A. Teixeira, L. Real, M. Garcia, H. G. Oliveira, and R. Amaro, Eds., Santiago de Compostela, Spain, Mar. 2024, pp. 321–331.
[53] L. Basyal and M. Sanghvi, "Text summarization using large language models: A comparative study of MPT-7b-instruct, falcon-7b-instruct, and OpenAI chat-GPT models," 2023, *arXiv:2310.10449*.
[54] H. Askari, A. Chhabra, M. Chen, and P. Mohapatra, "Assessing LLMs for zero-shot abstractive summarization through the lens of relevance paraphrasing," 2024, *arXiv:2406.03993*.
[55] C. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. Bethard, and D. McClosky, "The Stanford CoreNLP natural language processing toolkit," in *Proc. 52nd Annu. Meeting Assoc. Comput. Linguistics, Syst. Demonstrations*, 2014, pp. 55–60.
[56] M. Zhong, P. Liu, D. Wang, X. Qiu, and X. Huang, "Searching for effective neural extractive summarization: What works and what's next," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 1049–1058.
[57] R. Paulus, C. Xiong, and R. Socher, "A deep reinforced model for abstractive summarization," 2018, *arXiv:1705.04304*.
[58] M. Grootendorst, "BERTopic: Neural topic modeling with a class-based TF-IDF procedure," 2022, *arXiv:2203.05794*.
[59] D. Deutsch, R. Dror, and D. Roth, "A statistical analysis of summarization evaluation metrics using resampling methods," *Trans. Assoc. Comput. Linguistics*, vol. 9, pp. 1132–1146, Oct. 2021.
[60] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, and K. Macherey, "Google's neural machine translation system: Bridging the gap between human and machine translation," 2016, *arXiv:1609.08144*.
[61] 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. Conf. Empirical Methods Natural Lang. Process.*, 2018, pp. 1797–1807.
[62] I. Chamieh, T. Zesch, and K. Giebermann, "LLMs in short answer scoring: Limitations and promise of zero-shot and few-shot approaches," in *Proc. 19th Workshop Innov. Use NLP Building Educ. Appl.*, E. Kochmar, M. Bexte, J. Burstein, A. Horbach, R. Laarmann-Quante, A. Tack, V. Yaneva, and Z. Yuan, Eds., Mexico City, Mexico, Jun. 2024, pp. 309–315.
[63] A. Shah and S. Chava, "Zero is not hero yet: Benchmarking zero-shot performance of LLMs for financial tasks," 2023, *arXiv:2305.16633*.

KEERTHANA MURUGARAJ received the bachelor's degree in robotics and automation engineering from the PSG College of Technology, Coimbatore, India, in 2016, and the master's degree (Hons.) in information and computer science from the University of Luxembourg, in 2023. She was a Software Engineer with Accenture for two years. She was honored with the prestigious Germain Dondelinger Prize for her best master's thesis. Currently, she is a Ph.D. Researcher with the University of Luxembourg in the field of natural language processing.

SALIMA LAMSIYAH received the degree in computer science and mathematics and the master's degree in information systems, networks, and multimedia from the Faculty of Science Dhar El Mahraz, Sidi Mohamed Ben Abdellah University, in 2014 and 2016, respectively, and the Ph.D. degree in computer science from Ibn Tofail University, Morocco, in collaboration with Aix Marseille University, in December 2021. Currently, she is a Postdoctoral Researcher in natural language processing and machine learning with the University of Luxembourg. Her research interests include machine learning, natural language processing, and deep learning.

CHRISTOPH SCHOMMER received the degree in artificial intelligence from the University Saarbrücken and the Ph.D. degree in medical informatics from the Goethe University Frankfurt/Main.

He was with IBM for eight years. In October 2003, he was appointed as an Associate Professor with the University of Luxembourg. Currently, he heads the MINE Research Group and the ACC Laboratory. He regularly organizes lecture series/Ph.D. workshops and is the author of approximately 100 scientific papers. He supervised 30 Ph.D. projects in Luxembourg, Turin, and London, and oversees 12 Ph.D. projects. He has taught a total of 170 courses at the University of Luxembourg (computer sciences, mathematics, finance) and several other universities in the EU, Beijing (Tsinghua), and Singapore (SUTD). His research interests include artificial intelligence and the intersection of machine learning and data science. He is an internationally recognized scientific reviewer for the Leibniz Association, Springer, IEEE, and served as a PC member at more than 100 international conferences (such as IJCAI, AAMAS, CogSci, and ECML). He maintains contacts with industry and the National Ethics Council. He is a member of the ACM, the Cognitive Science Society, and the Deutsche Gesellschaft für Kognitionswissenschaft.