

RESEARCH TRENDS IN TEXT SUMMARIZATION WITH A FOCUS ON FINE-TUNING INSTRUCTION-TUNED LANGUAGE MODELS

**Thesis Submitted
in Partial Fulfillment of the Requirements for the
Degree of**

**MASTER OF TECHNOLOGY
in
Computer Science And Engineering
by**

**Rishabh Chauhan
(Roll No. 23/CSE/13)**

**Under The Supervision Of
Ms. ANUKRITI KAUSHAL
Assistant Professor, Computer Science and Engineering
Delhi Technological University**



Department of Computer Science and Engineering

**DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Shahbad Daultpur, Main Bawana Road, Delhi-110042, India**

MAY, 2025



DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Shahbad Daulatpur, Main Bawana Road, Delhi-42

CANDIDATE'S DECLARATION

I, Rishabh Chauhan, 23/CSE/13 of M.Tech (Computer Science and Engineering), hereby certify that the work which is being presented in the thesis entitled "Research Trends In Text Summarization With A Focus On Fine-Tuning Instruction-Tuned Language Models" in partial fulfillment of the requirements for the award of the degree of Master of Technology, submitted in the Department of Computer Science and Engineering, Delhi Technological University is an authentic record of my own work carried out during the period from to under the supervision of Ms. Anukriti Kaushal.

The matter presented in the thesis has not been submitted by me for the award of any other degree of this or any other institute.

Candidate's Signature

This is to certify that the student has incorporated all the corrections suggested by the examiners in the thesis and the statement made by the candidate is correct to the best of our knowledge.

Signature of Supervisor

Signature of External Examiner



DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Shahbad Daultpur, Main Bawana Road, Delhi-42

CERTIFICATE BY THE SUPERVISOR

Certified that **Rishabh Chauhan**(23/CSE/13) has carried out their search work presented in this thesis entitled **"Research Trends In Text Summarization With A Focus On Fine-Tuning Instruction-Tuned Language Models"** for the award of **Master of Technology** from Department of Computer Science and Engineering, Delhi Technological University, Delhi, under my supervision. The thesis embodies results of original work, and studies are carried out by the student himself and the contents of the thesis do not form the basis for the award of any other degree to the candidate or to anybody else from this or any other University/Institution.

Signature

(Ms. Anukriti Kaushal)

(Assistant Professor)

(DELHI TECHNOLOGICAL UNIVERSITY)

(Formerly Delhi College of Engineering)

Shahbad Daultpur, Main Bawana Road, Delhi-42)

Date: 30 May 2025

RESEARCH TRENDS IN TEXT SUMMARIZATION WITH A FOCUS ON FINE-TUNING INSTRUCTION-TUNED LANGUAGE MODELS

RISHABH CHAUHAN

ABSTRACT

Text summarisation has emerged as a crucial area within Natural Language Processing, driven by the increasing need to process and condense vast amounts of textual data. This thesis presents a twofold investigation into the domain: (1) a comprehensive bibliometric analysis of global research trends in text summarisation, and (2) a performance-focused study on hyperparameter tuning for domain-specific text generation using transformer-based models.

The first study employs bibliometric methods, visual analytics, and critical review to trace the evolution of text summarisation across decades, highlighting key contributors, publication trends, influential methodologies, and emerging directions such as multilingual and multimodal summarisation. This exploration not only contextualizes the field's growth but also identifies research gaps and strategic opportunities for innovation.

Building upon these insights, the second study investigates the practical challenges of adapting instruction-tuned models for domain-specific summarisation. Using the BillSum legal dataset, the performance of two models—BART-Large and FLAN-T5 Base—was evaluated. Initial experiments revealed that FLAN-T5 underperformed significantly compared to BART-Large. Traditional strategies such as data augmentation and cross-domain pretraining failed to yield improvements. However, targeted hyperparameter optimization led to substantial gains, emphasizing the critical role of configuration over model architecture alone.

Together, these studies offer both a macro and micro perspective of the field: from understanding its research trajectory to demonstrating how careful tuning can unlock the potential of underperforming models for specialized summarisation tasks. The findings aim to inform both researchers and practitioners seeking to advance summarisation technologies in theory and application.

ACKNOWLEDGEMENTS

I wish to express my sincerest gratitude to Ms. Anukriti Kaushal for her continuous guidance and mentor-ship that she provided during the M.Tech thesis. She showed me the path to achieving targets by explaining all the tasks to be done and explained the importance of this work as well as its industrial relevance. She was always ready to help and clear doubts regarding any hurdles in this project work. Without her constant support and motivation, this project would not have been successful.

(RISHABH CHAUHAN)

CONTENTS

Title	Page No.
Candidate's Declaration	II
Certificate	III
Abstract	IV
Acknowledgements	V
Contents	VI
List of Tables	VIII
List of Figures	IX
List of Abbreviations	XI
CHAPTER 1: INTRODUCTION	1
1.1 Overview	1
1.2 Evolution of Text Summarisation Techniques	1
1.3 Motivation for Bibliometric Exploration	2
1.4 Challenges in Domain-Specific Summarisation	2
1.5 Research Objectives	2
CHAPTER 2: RELATED WORK	3
2.1 Literature review	3
2.2 Instruction Tuned Models	3
2.3 Denoising Language Models	4
2.4 Datasets	6
CHAPTER 3: METHODOLOGY	9

3.1 Model Optimization Methodology	9
3.2 Combined Analysis of Publication Trends, Citations, and Collaboration Networks	10
3.3 Citation and Collaboration Metrics	13
3.4 Co-Authorship and Co-Citation Networks	16
3.5 Publisher Landscape	17
CHAPTER 4: RESULT	18
4.1 Experimental Results on Model Performance	18
CHAPTER 5: CONCLUSION, FUTURE SCOPE, SOCIAL IMPACT	26
5.1 Conclusion	26
5.2 Future Scope	27
5.3 Social Impact	27
REFERENCES	28
LIST OF PUBLICATIONS	30

LIST OF TABLES

Table No.	Table Name	Page No.
Table 3.4.1	Top researchers with the highest TLS ratings.	16
Table 4.1.1	Bart-Large On Billsum With 8000 Rows In Train Set And 1000 Rows In Test Set	18
Table 4.1.2	FLAN-T5-Base on BillSum with 8000 Rows in Train Set and 1000 Rows in Test Set	20
Table 4.1.3	FLAN-T5-Base on BillSum, With 8000 Rows in Train Set and 1000 Rows in Test Set, After Training on XSum	21
Table 4.1.4	Flan-T5-Base on BillSum With 18.9K Rows in Training and 3.27K Rows in Test Set.	23
Table 4.1.5	FLAN-T5-Base on BillSum After Hyperparameter Tuning	24

LIST OF FIGURES

Figure No.	Figure Name	Page No.
Fig 2.3.1	An overview of Text Summarisation Techniques.	5
Fig 2.4.1	Prisma diagram for search query.	7
Fig 3.2.1.1	An annual publication growth graph illustrating the rise in publications from 2015 to 2024.	11
Fig 3.2.1.2	The top 10 categories in the publications.	12
Fig 3.2.2.1	The top ten research directions in the articles indicate the main areas of attention.	13
Fig 3.3.1	The ten most productive locations.	14
Fig 3.3.2	The organizations that produce the greatest number of research publications.	15
Fig 3.3.3	The collaborative network of places.	15
Fig 3.4.1	The network of co-citations for articles.	16
Fig 3.5.1	Chord diagram of publisher and publishing year.	17
Fig 4.1.1	Loss graph for Table 4.1.1	19
Fig 4.1.2	Rouge score graph for Table 4.1.1	19
Fig 4.1.3	Loss graph for Table 4.1.2	20
Fig 4.1.4	Rouge score graph for Table 4.1.2	21
Fig 4.1.5	Loss graph for Table 4.1.3	22
Fig 4.1.6	Rouge score graph for Table 4.1.3	22
Fig 4.1.7	Loss graph for Table 4.1.4	23

Fig 4.1.8	Rouge score graph for Table 4.1.4	24
Fig 4.1.9	Loss graph for Table 4.1.5	25
Fig 4.1.10	Rouge score graph for Table 4.1.5	25

LIST OF ABBREVIATIONS

Abbreviation	Meaning
TF-IDF	Term Frequency-Inverse Document Frequency
FLAN-T5	Fine-Tuned Language Model - T5
GPT	Generative Pre-training Transformer
BERTSUM	Variant of BERT for extractive summarisation
BART	Bidirectional and Auto-Regressive Transformers
NLP	Natural Language Processing
ROUGE	Recall-Oriented Understudy for Gisting Evaluation
SCI-EXPANDED	Science Citation Index Expanded
SSCI	Social Sciences Citation Index
A&HCI	Arts and Humanities Citation Index
CPCI-S	Conference Proceedings Citation Index – Science

CHAPTER 1

INTRODUCTION

1.1 Overview

Within the field of Natural Language Processing (NLP), automatic text summarization has emerged as a crucial field that provides effective means of gleaning pertinent information from vast amounts of data. The skill to write coherent and short summaries has applications across a variety of sectors, including healthcare, education, media, and especially law—where lengthy papers and thick textual material are widespread. This thesis explores two complementary aspects of this field: the evolutionary landscape of text summarisation research and the optimization of modern summarisation models for domain-specific tasks.

1.2 Evolution of Text Summarisation Techniques

Text summarisation approaches have evolved substantially since their inception. Early summarisation methods in the 1990s relied on extractive techniques using statistical and frequency-based scoring methods such as TF-IDF. These models as often as possible delivered outlines that needed consistency and relevant stream since they chose key sentences from the content verbatim. Improved significance estimation and excess decrease were created by machine learning models and graph-based strategies like TextRank to overcome these constraints.

Abstractive summarization, in which models create rundowns by summarizing instead of reproducing the crude content, was made conceivable by the presentation of transformer-based structures, which totally changed the teach. BERTSUM[11], PEGASUS, GPT-2/3/4, and instruction-tuned[4] models like FLAN-T5 are striking turning focuses in this movement. These models significantly moved forward the created summaries' coherence, familiarity, and semantic understanding, which qualified them for down to earth utilization.

1.3 Motivation for Bibliometric Exploration

Given the fast development of summarization investigation, it is basic to chart its advancement all through time. A bibliometric examination offers a deliberate way to discover critical mechanical advancements, creating subjects, and striking donors. This proposition offers a data-driven point of view on the advancement of summarization investigate by applying devices like VOSviewer and RawGraphs to information accumulated from dependable databases like Web of Science. In expansion to recognizing authentic designs, this approach recognizes critical roads to encourage inquiry about.

1.4 Challenges in Domain-Specific Summarisation

Indeed with the progressions in demonstrate models, execution shifts by domain. The complexity of dialect structures, domain-specific lexicon, and the prerequisite for tall truthful precision make summarizing legitimate records especially troublesome. Due to complicated state developments, domain-specific lexicon, and the necessity for tall real precision, lawful record summarization presents special difficulties. Important inquiries concerning demonstrate generalizability and fine-tuning methods are brought up by this execution dissimilarity. Expanding the sum of the dataset or utilizing cross-domain pretraining are cases of routine procedures that do not continuously deliver picks up. Or maybe, there's mounting prove that making strides demonstrate execution for certain assignments requires hyperparameter alteration.

1.5 Research Objectives

Two roads of request are combined in this thesis:

1. To give a basic and bibliometric examination of content summarization investigate, spotting patterns, holes, and modern roads for the discipline.
2. Through deliberate hyperparameter tuning, the FLAN-T5[8] model's execution on the BillSum[10] dataset will be surveyed and made strides in comparison to BART-Large's pattern performance.

This proposal gives an in-depth understanding of the macro-level headway and micro-level challenges in content summarization by combining slant investigation with valuable demonstration optimization.

CHAPTER 2

RELATED WORK

2.1 Literature review

Ordinary studies, such as those conducted by El-Kassas et al. (2021)[16] and Gambhir & Gupta (2017)[17], concentrated for the most part on mechanical improvements in summarization calculations. They did not, be that as it may, give a quantitative appraisal of the state of the field. Later investigate has filled this vacuum by examining distribution patterns, co-authorship systems, and topical clusters in summarization considers utilizing bibliometric instruments like VOSviewer and RawGraphs.

By looking at nearly a thousand distributions from the Net of Science database, this proposal extends on earlier ponders and gives a data-driven and visual picture of the field's advancement. The investigation centers on modern themes counting domain-specific summarization, real-time summarization, and moral concerns like preference and visualization in expansion to more customary subjects like extractive and abstractive models. Master assessment and bibliometric investigation work together to supply a well-organized investigation guide.

2.2 Instruction Tuned Models

A powerful strategy for improving show generalization over NLP assignments is instruction tuning. Instruction tuning trains models on a huge collection of instruction-response pairings, empowering zero-shot and few-shot capabilities, in differentiate to conventional fine-tuning, which customizes a show for a specific work.

One well-known outline of an instruction-tuned transformer that exceeds expectations at summarizing, interpreting, and replying questions is the FLAN-T5 type. Such models perform conflictingly in domain-specific settings, such lawful or therapeutic summarization, however exceeding expectations at non specific errands.

Typically primarily since, especially when domain-specific semantics and dialect are included, instruction tweaking might not offer sufficient task-specific presentation.

2.3 Denoising Language Models

Denoising autoencoders are utilized in pre-training for models such as BART (Bidirectional and Auto-Regressive Transformer). Solid semantic understanding and creation capabilities are made conceivable by their capacity to reproduce the first content from mutilated input. Since BART employs autoregressive decoders and bidirectional encoders to reason over the whole input arrangement, it has beaten various pattern models in abstractive summarization.

Mental trips, on the other hand, can moreover cause denoising models to create content that's liquid but genuinely wrong. In high-stakes areas where truthful precision is significant, like lawful and restorative summarization, this gets to be a genuine concern. Hence, when utilizing such models for domain-specific errands, exhaustive evaluation and alteration are significant.

A careful survey of content summarization strategies is given in Figure 2.3.1, where they are separated into two essential categories: extractive and abstractive. In extractive outline, vital articulations or expressions are chosen straight from the first content, without any rewording. This category incorporates factual methods that gauge sentence pertinence utilizing word recurrence and graph-based strategies like TF-IDF, TextRank[18], and LexRank[19]. To decide which sections of the content to extricate, machine learning procedures for extractive summarization utilize classifiers like Gullible Bayes, Choice Trees, and SVMs in expansion to neural network-based models like RNNs, CNNs, and transformers.

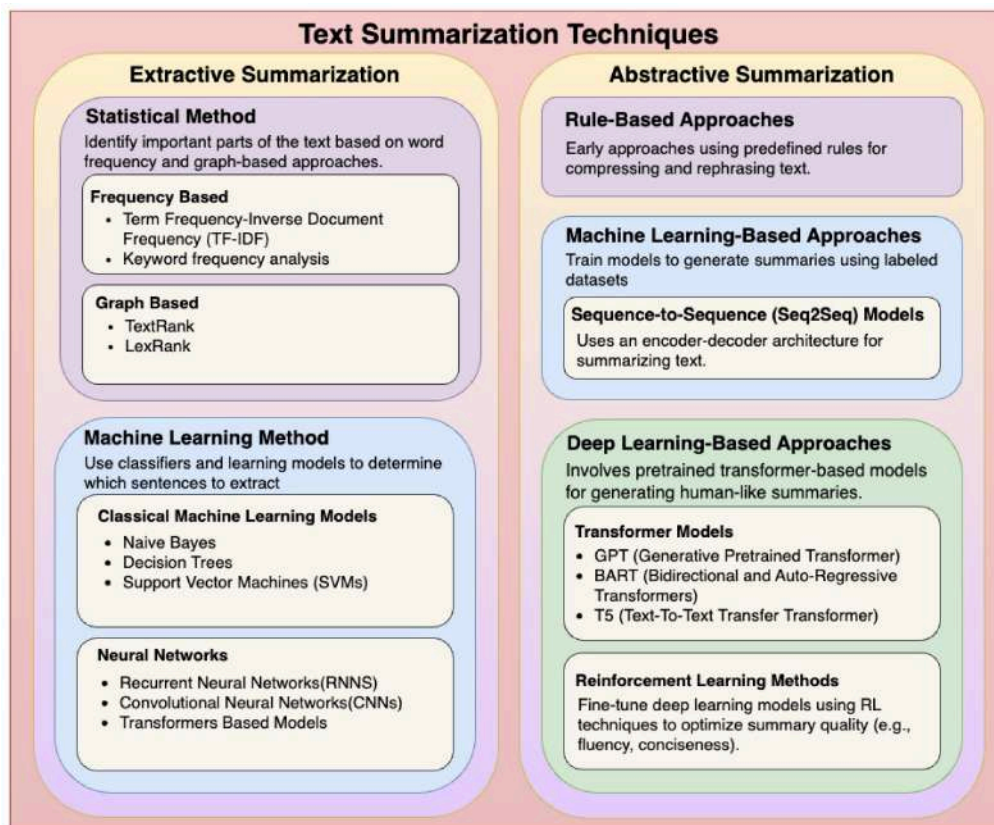


Figure 2.3.1 - An overview of Text Summarisation Techniques.

On the other hand, abstractive summarizing, which regularly requires substance creation and rethinking, produces modern sentences that express the most concepts of the first fabric. Rule-based strategies that utilize pre-established phonetic rules for content modifying and compression drop beneath this category. More advanced strategies depend on machine learning, particularly encoder-decoder models for sequence-to-sequence (Seq2Seq) models. By utilizing powerful pretrained transformer models like GPT, BART, and T5, which are planning to supply consistent and relevantly fitting outlines, profound learning-based methods encourage progress summarization capabilities. These models are also improved through the use of reinforcement learning techniques, which maximize output quality in terms of conciseness and fluency. All things considered, the picture depicts the variety and development of summarization techniques, from basic statistical models to complex neural and transformer-based systems.

2.4 Datasets

Two different kinds of datasets, each representing one of the research components, are used in this thesis: (1) text summarization corpora, particularly XSum and BillSum, for assessing and improving model performance in domain-specific summarization, and (2) a scientific publishing dataset from bibliographic databases for bibliometric study of summarization trends.

The Web of Science (WoS) Core Collection, a well-known bibliographic database renowned for its well curated and superior indexing of scientific articles, provided the data used for the trend analysis of text summarization research. The Core Collection includes the following citation indexes:

- SCI-EXPANDED (Science Citation Index Expanded)
- SSCI (Social Sciences Citation Index)
- A&HCI (Arts and Humanities Citation Index)
- CPCI-S (Conference Proceedings Citation Index – Science)

A carefully formulated topic search query was used:

TS = (“Text summarization” OR “text summarizes” OR “text summarizing”
OR “Text summarisation” OR “text summarises” OR “text summarising”)

This query was designed to capture all common and variant spellings of the term “text summarisation” across titles, abstracts, author keywords, and keywords plus. Results were filtered using a PRISMA flow (see Figure 2.4.1) to ensure precision and eliminate duplicates. WoS's unique identifier system (UT) helped maintain data integrity and ensured each publication was counted only once.

The final dataset included:

- 983 articles published from 2004 to September 2024
- 18,392 total citations (14,833 excluding self-citations)
- 338 journals, 2952 authors, and a collective H-index of 60

- All documents were in English, filtered by type to emphasize original research and scholarly reviews

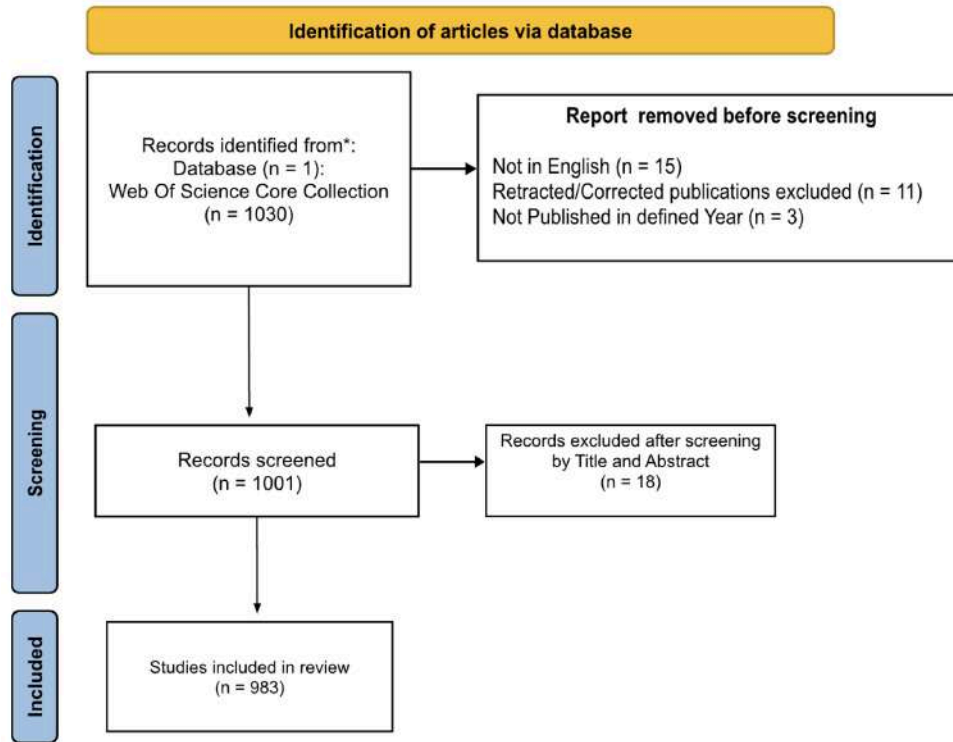


Fig. 2.4.1 - Prisma diagram for search query (the number of records found from each database searched is shown by the symbol "*").

This dataset provided the foundation for identifying thematic evolution, co-authorship networks, and emerging subfields in text summarisation.

To investigate model behavior in domain-specific summarisation, two prominent benchmark datasets were used:

The XSum dataset, created by the BBC, comprises single-sentence news summaries for over 200,000 articles. It was used to explore whether pretraining on highly abstractive summarisation data could enhance generalisation for other tasks. Each summary in XSum is designed to be maximally informative, often rephrasing the content entirely, thus making it well-suited for training generative models like FLAN-T5 and BART.

BillSum is a domain-specific dataset that includes human-written summaries of 22,000 California state and U.S. congressional bills. It illustrates how difficult it is to summarize legal papers due to their language and structural complexity. Due to the formal, technical language and requirement for factual correctness, BillSum serves as a rigorous benchmark for evaluating summarisation model adaptability in legal contexts.

Initial training and evaluation were performed on BillSum, while additional fine-tuning on XSum was explored to test whether cross-domain pretraining could enhance model performance. Despite the high quality of both datasets, performance improvements were realized only after targeted hyperparameter tuning, highlighting the need for tailored configurations in domain-specific applications.

CHAPTER 3

METHODOLOGY

3.1 Model Optimization Methodology

3.1.1 Model Selection

Two transformer-based models from the HuggingFace library were selected:
Model 1: FLAN-T5 Base (Instruction-tuned)
Model 2: BART-Large (Denoising pre-trained)
These models were chosen to evaluate performance disparities on legal-domain summarisation and to explore improvement strategies for underperforming models.

3.1.2 Dataset Description

The following summarisation datasets were used:
BillSum: A legal summarisation corpus of U.S. congressional bills and their summaries.
XSum: A highly abstractive summarisation dataset of BBC news articles.

3.1.3 Experimental Workflow

Step 1: Baseline Evaluation

The BillSum dataset was used to refine both models.
ROUGE-1 scores were computed to establish baseline performance.
BART-Large: 40.09
FLAN-T5 Base: 18.39

Step 2: Dataset Augmentation

Hypothesis: More training data improves performance.
Two attempts:
Pretraining FLAN-T5 on XSum followed by BillSum fine-tuning.
Training on an expanded BillSum dataset.

Outcome: Marginal or no performance gain (ROUGE[1] remained ~18–25).

Step 3: Hyperparameter Optimization

Hypothesis: Adjusting generation parameters would improve summarisation quality.

Parameters tuned:

num_beams = 6

generation_max_length = 300

Result:

FLAN-T5 Base's ROUGE-1 improved to 40.85, nearly matching BART-Large.

3.1.4 Evaluation Metrics

Training Loss and Validation Loss monitored per epoch. ROUGE-1 used for quantifying summary overlap between predictions and ground truth.

Seaborn[9] visualizations were created to illustrate performance changes.

3.2 Combined Analysis of Publication Trends, Citations, and Collaboration Networks

This section presents a consolidated view of publication structures, citation patterns, institutional impact, and publisher contributions in text summarisation research.

3.2.1 Annual Publication Growth and Document Types

Between 2004 and 2024, there has been steady growth in text summarisation research, culminating in a total of 983 publications. The majority (95.7%) are research articles, followed by reviews and conference papers. This indicates a strong foundation of peer-reviewed contributions to the field(Figure 3.2.1.1 and 3.2.1.2).

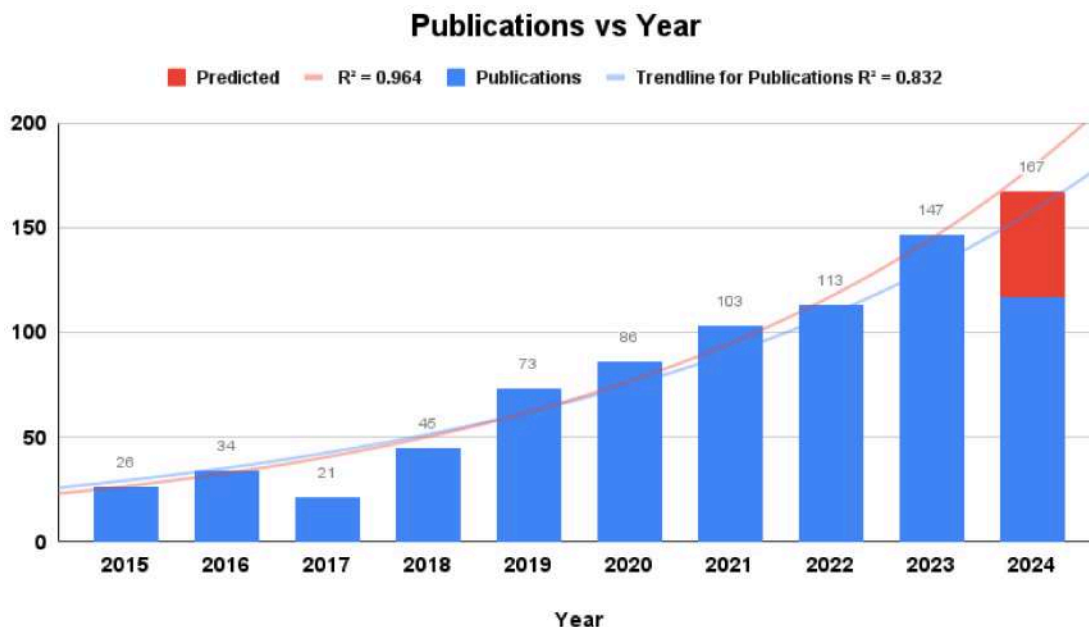


Fig. 3.2.1.1 - An annual publication growth graph illustrating the rise in publications from 2015 to 2024.

A thorough examination of the publishing categories of the 983 documents that were obtained from the Web of Science (WoS) database is shown in Figure 3.2.1.2. The data reveals that research articles constitute the vast majority, with 941 entries, accounting for approximately 95.72% of the total publications. This dominance highlights the field's strong focus on original research contributions. In addition to research articles, the dataset includes 41 review articles (4.17%), which provide comprehensive overviews of existing literature. There are also 25 early access papers (2.54%) and 40 conference proceedings (4.07%), reflecting the dissemination of ongoing and preliminary findings. A smaller proportion is made up of letters and other miscellaneous types, indicating a diverse but research-centric publication landscape.

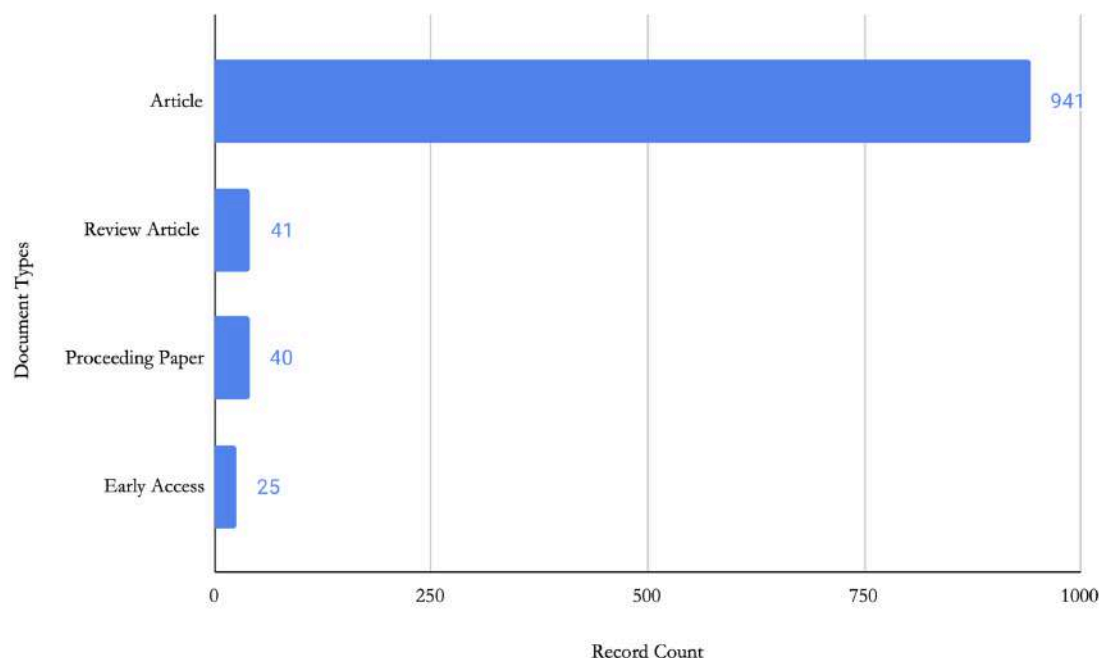


Fig. 3.2.1.2 - The top 10 categories in the publications.

3.2.2 Disciplinary and Geographic Trends

Electrical engineering, information systems, and artificial intelligence are the most active WoS categories. Geographically, the leading contributors are the USA (129), India (172), and China (215 publications). Among the most prominent are the Chinese Academy of Sciences and the Indian Institute of Technology (IIT) system (Figure 3.2.2.1).

Fig.-3.2.2.1 highlights the significant impact of various academic domains on the field of text summarisation. Leading the contributions is the field of Computer Science, particularly the subdomain of Artificial Intelligence, which accounts for 382 publications—approximately 38.86% of the total. A more thorough analysis reveals that the number of publications in computer science not only surpasses but also might overlap by about 955 articles with those in electrical and electronic engineering. Artificial Intelligence, Theory and Methods, Information Systems, Software Engineering, and Multidisciplinary Applications are some of the subfields in computer science that contribute to the literature. The interdisciplinary character and broad applicability of summarization research are further demonstrated by the 80 papers in telecommunications, 46 in multidisciplinary engineering, 34 in medical informatics, and 27 in linguistics.

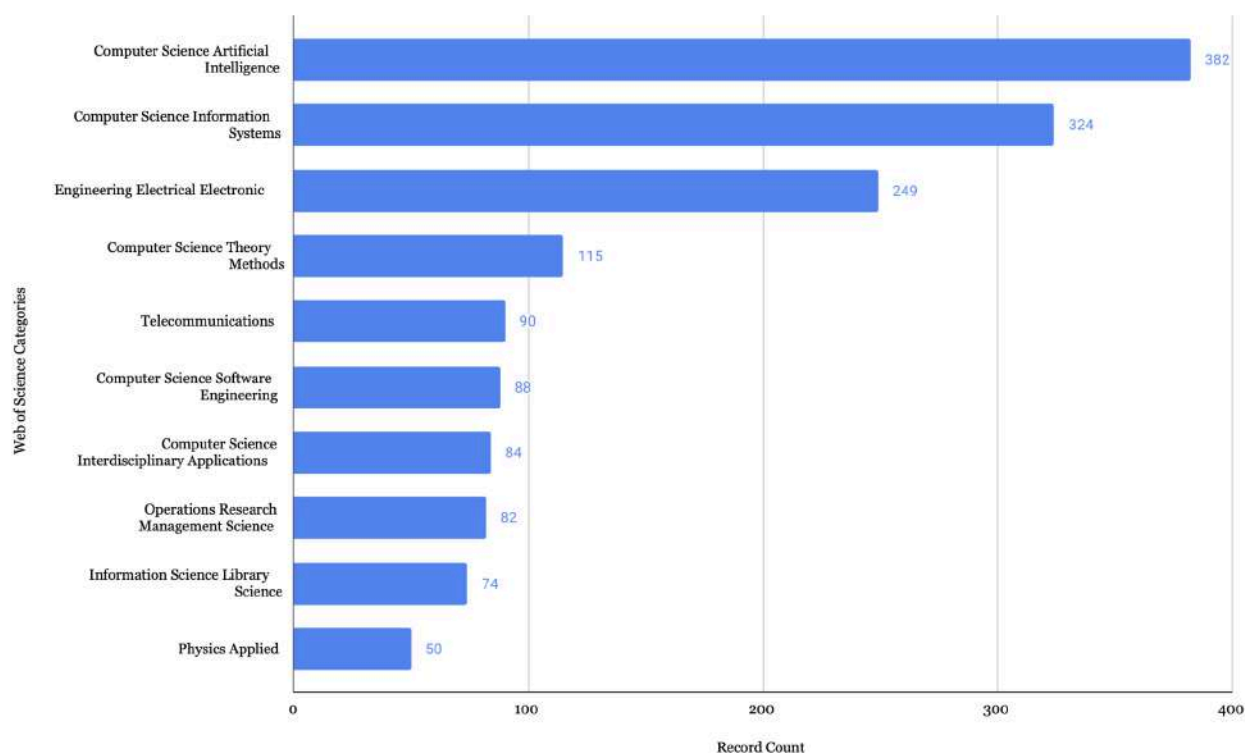


Fig. 3.2.2.1 - The top ten research directions in the articles indicate the main areas of attention.

3.3 Citation and Collaboration Metrics

According to citation analysis, the countries with the highest average number of citations per manuscript are the USA, China, and Italy. Institutions with a high citation strength and collaborative participation include Jawaharlal Nehru University and the University of Chinese Academy of Sciences. Prominent authors include Lloret E., Palomar M., and Devlin J(Figure 3.3.1, 3.3.2 and 3.3.3).

Fig.-3.3.2 illustrates the leading institutional contributors to text summarisation research based on publication count. With 31 publications, the Indian Institute of Technology (IIT) System comes in first, demonstrating its substantial contribution to the subject. Following closely behind are the National Institute of Technology (NIT) System with 28 publications and the Chinese Academy of Sciences with 29, demonstrating China and India as significant centers of research. The Egyptian Knowledge Bank (EKB) also demonstrates notable engagement, contributing 22 publications. Other institutions with substantial outputs include the University of Chinese Academy of Sciences (16), King Saud University and Universitat d'Alacant (15 each), the Institute of Computing Technology CAS (11), and both Islamic Azad

University and Universidad de Extremadura, each with 10 publications. Overall, the chart reflects a strong international presence in text summarisation research, with particularly high contributions from institutions in India, China, the Middle East, and parts of Europe.

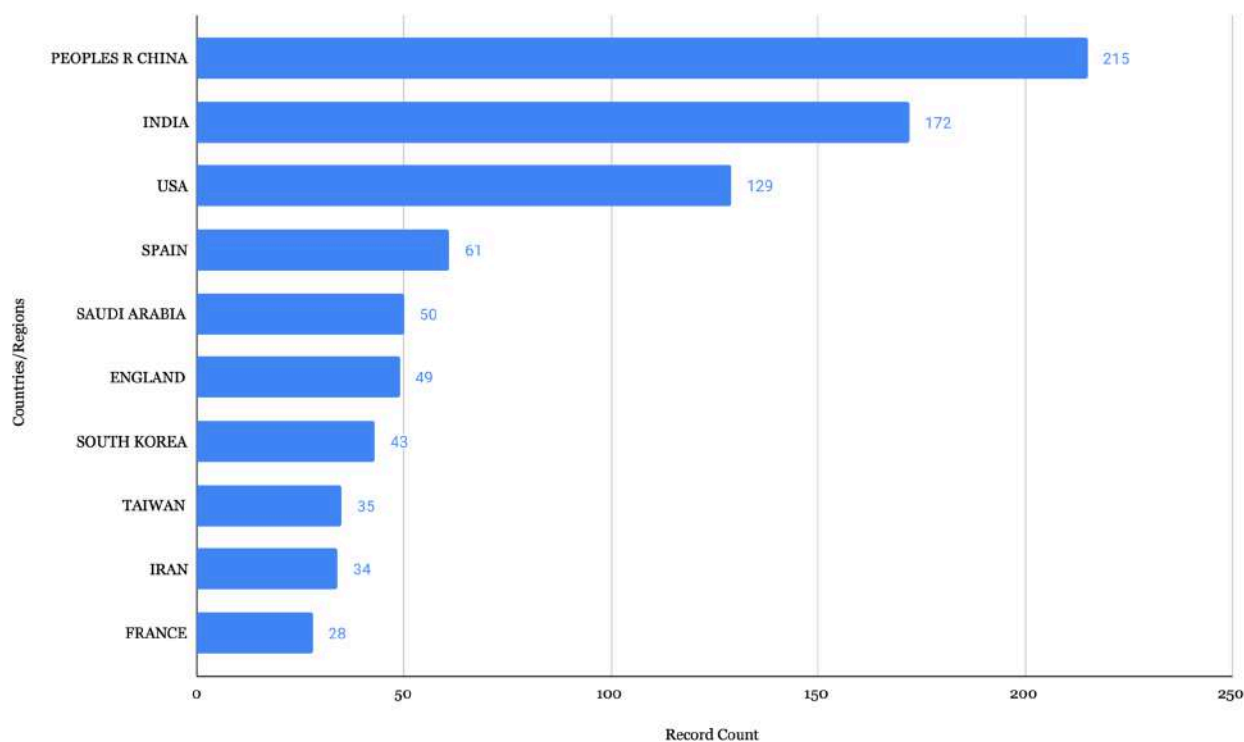


Figure 3.3.1 - The ten most productive locations.

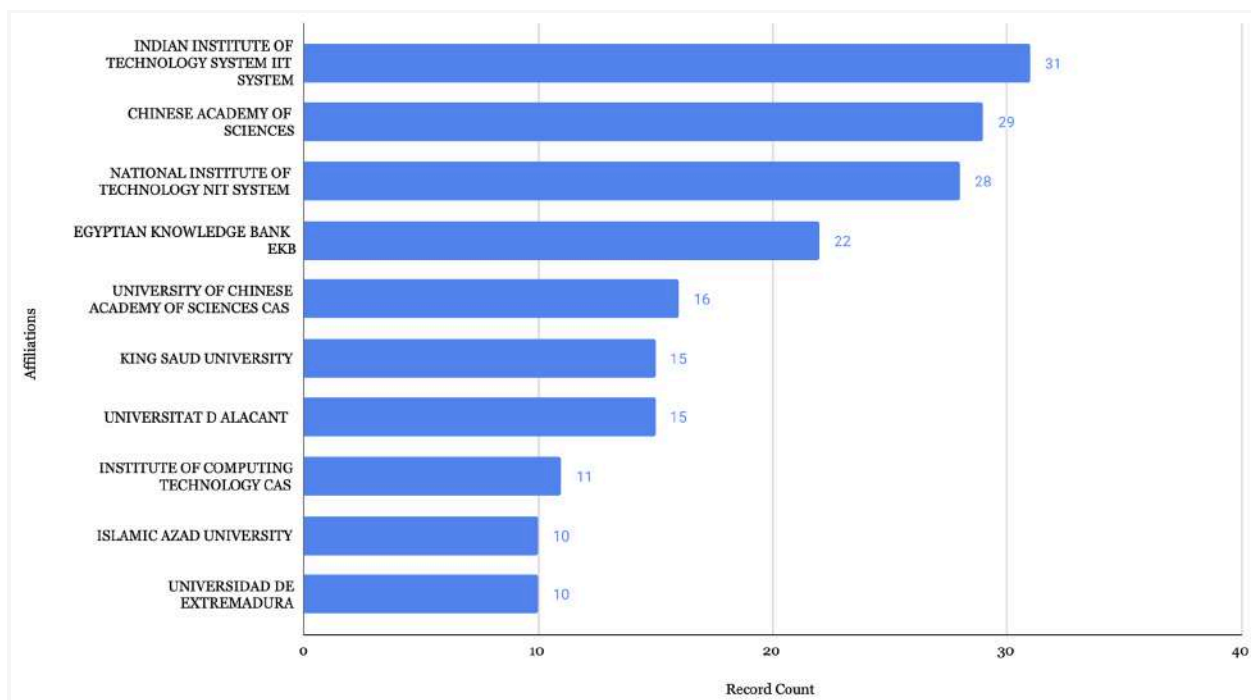


Figure 3.3.2 - The organizations that produce the greatest number of research publications.

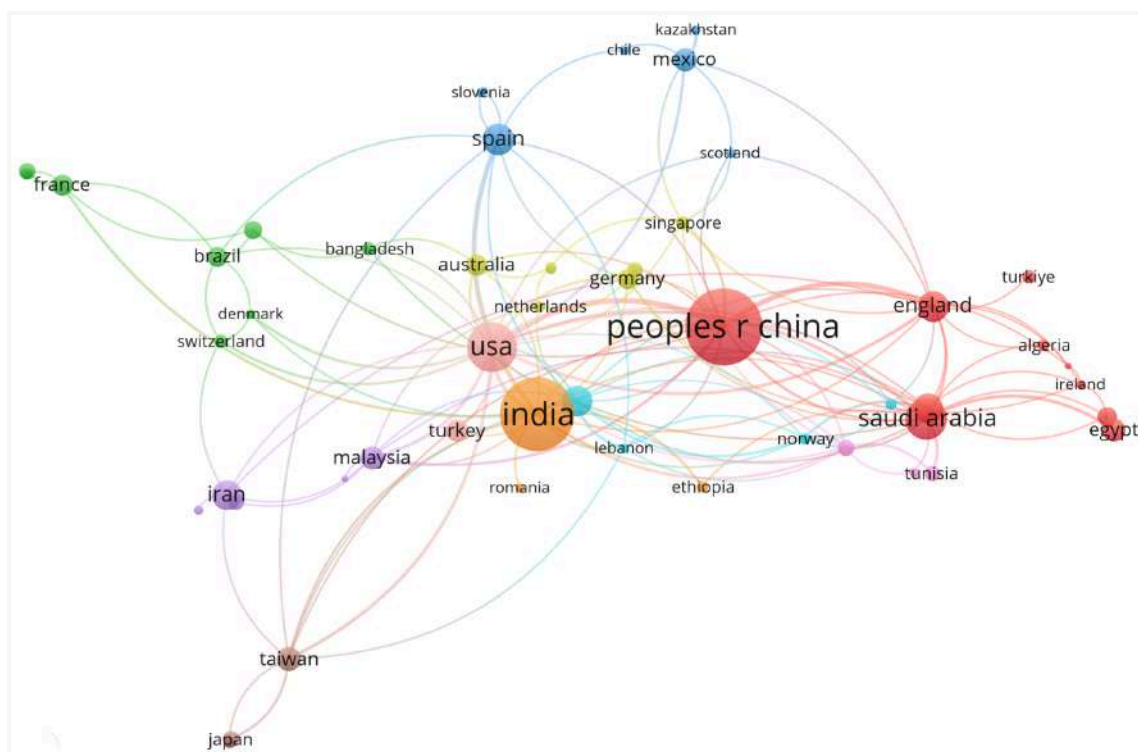


Figure 3.3.3 - The collaborative network of places.

3.4 Co-Authorship and Co-Citation Networks

Table 3.4.1 - Top researchers with the highest TLS ratings.

Rank	Authors	TLS	C
1	Nallapati, R	1565	252
2	Lin, CY	1513	284
3	See, A	1144	187
4	Mihalcea, R	953	154

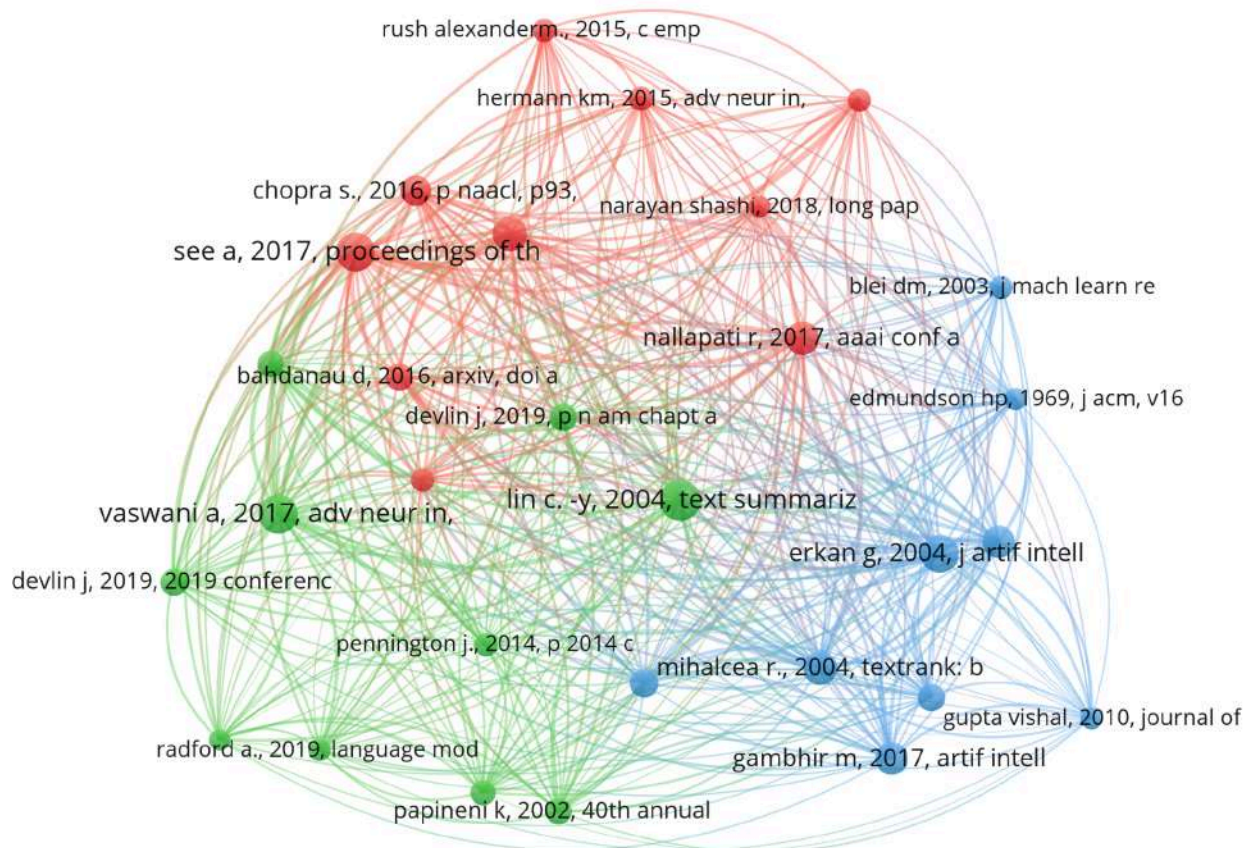


Fig. 3.4.1 - The network of co-citations for articles.

Using VOSviewer, co-authorship networks highlight clusters of collaboration between researchers and institutions. Scholars like Nallapati R. and Lin C.Y. exhibit

strong co-citation metrics. Influential papers include “Attention Is All You Need” (Vaswani et al.) and “BERT: Pre-training of Deep Bidirectional Transformers” (Devlin et al.).

3.5 Publisher Landscape

Chord diagrams generated using RawGraphs show that Springer, Elsevier, and IEEE have been the most active publishers from 2010 to 2024. Emerging contributors include MDPI and PeerJ Inc., indicating a diversification in publication venues.

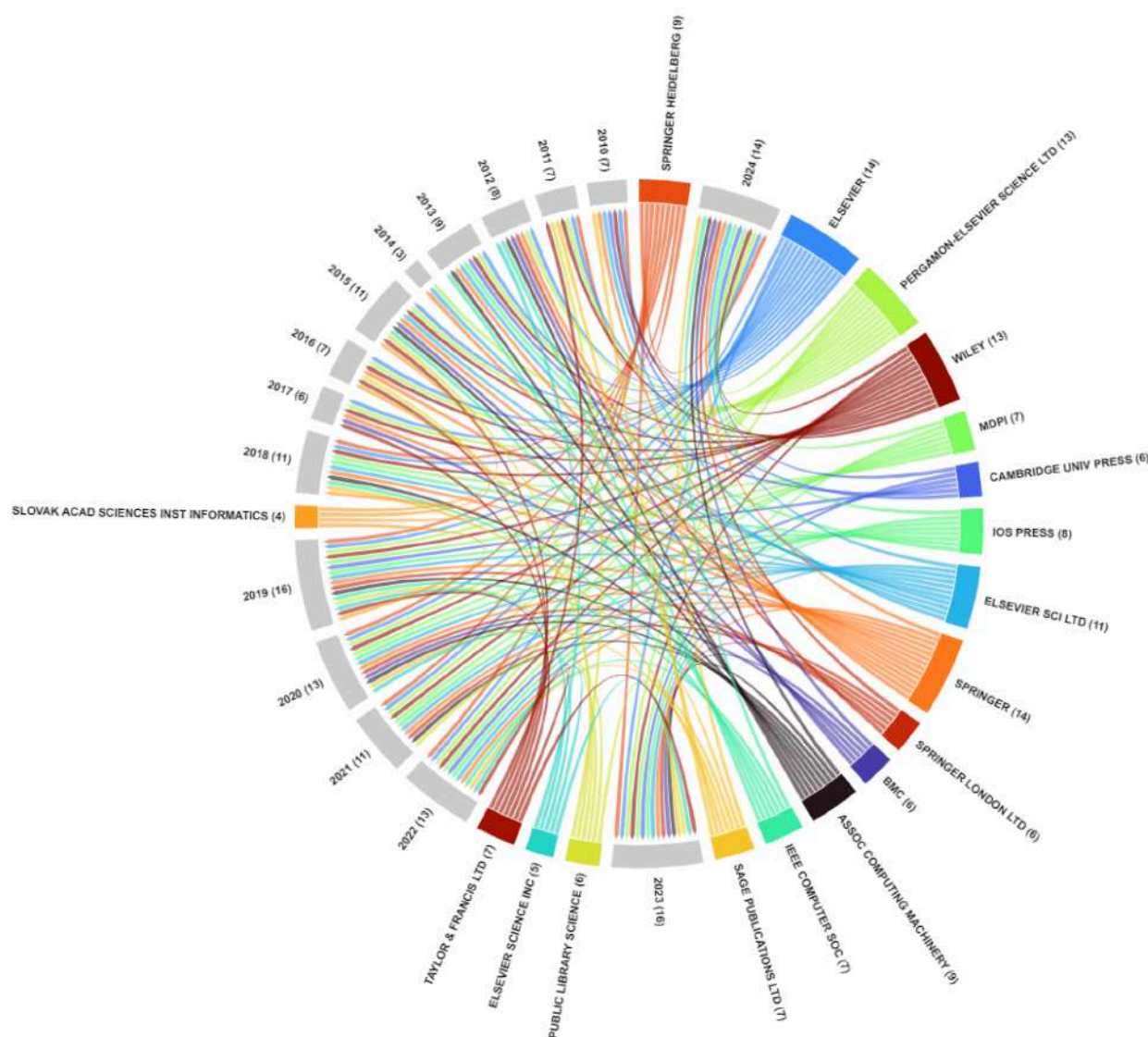


Fig. 3.5.1 - Chord diagram of publisher and publishing year.

CHAPTER 4

RESULTS

4.1 Experimental Results on Model Performance

This chapter presents tables and graphs illustrating ROUGE scores, training loss, and validation loss metrics based on the improvement strategies outlined in the methodology section. All visualizations were generated using the Seaborn Python library. Below is a brief explanation of the columns included in the tables.

1. Epoch: Shows the number of training cycles, with each epoch representing a full run of the training dataset.
2. Training Loss: Shows the error of the model on the training set; smaller values indicate a better match.
3. Validation Loss: Evaluates the model's capacity for generalization by reflecting its inaccuracy on unseen validation data.
4. ROUGE-1: Determines the content similarity between generated and reference summaries by measuring the unigram (single-word) overlap.

In Table 4.1.1, With training loss dropping from 1.4 to 1.08 and validation loss down from 1.4 to 1.3, BART-Large exhibits good performance. Additionally, the ROUGE-1 score somewhat improves, going from 39.8 to 40.0.

Table 4.1.1 - Bart-Large On Billsum With 8000 Rows In Train Set And 1000 Rows In Test Set

Epoch	Training Loss	Validation Loss	Rouge1
1	1.4711	1.430512	39.8935
2	1.2866	1.375137	39.9252
3	1.1315	1.372435	39.9512
4	1.0851	1.367797	40.0978

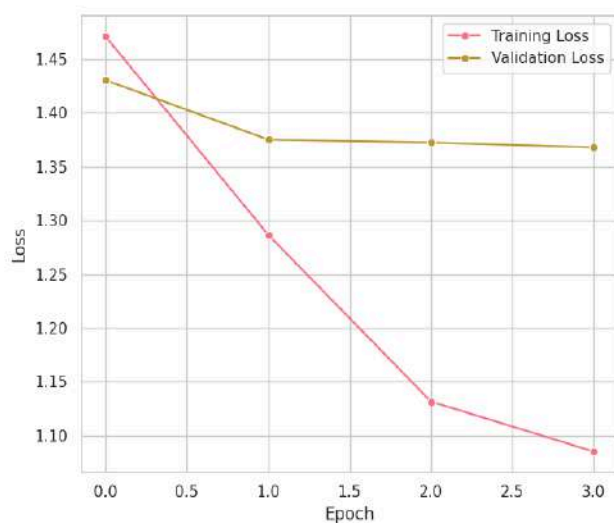


Fig. 4.1.1 - Loss graph for Table 4.1.1

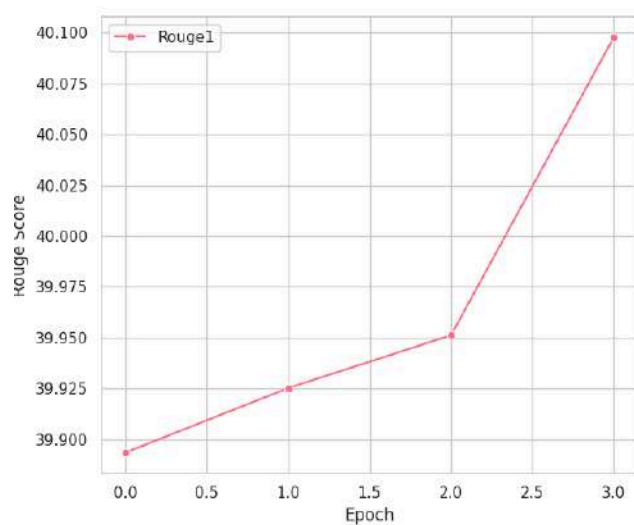


Fig. 4.1.2 - Rouge Score for Table 4.1.1

Table 4.1.2 demonstrates that FLAN-T5 performs worse than BART-Large. The validation loss falls from 1.7 to 1.6, and the training loss falls from 1.8 to 1.6. A slight improvement is shown in the ROUGE-1 score, which rises from 18.2 to 18.3.

Table 4.1.2 - FLAN-T5-Base on BillSum with 8000 Rows in Train Set and 1000 Rows in Test Set

<i>Epoch</i>	<i>Training Loss</i>	<i>Validation Loss</i>	<i>Rouge1</i>
1	1.8348	1.740196	18.2941
2	1.8827	1.691753	18.2926
3	1.6701	1.674098	18.4369
4	1.65	1.670359	18.3968

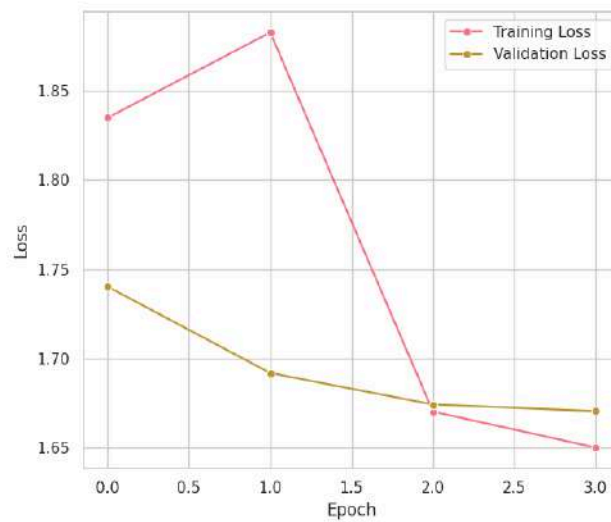


Fig. 4.1.3 - Loss graph for Table 4.1.2

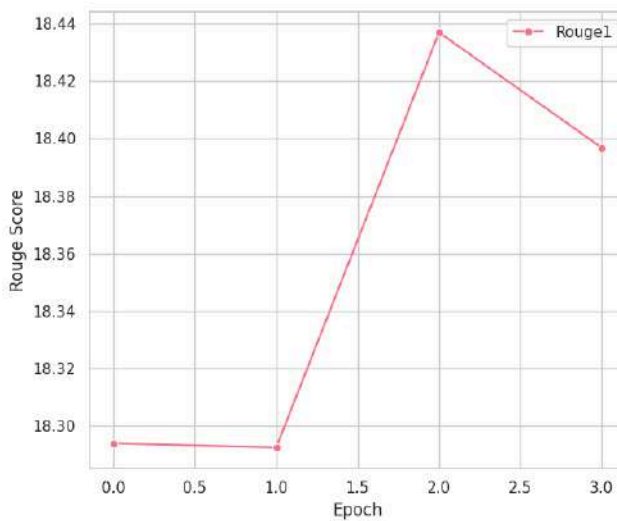


Fig. 4.1.4 - Rouge score graph for Table 4.1.2

In Table 4.1.3, FLAN-T5 shows a slight improvement in performance. The validation loss falls from 1.19 to 1.17, and the training loss somewhat decreases from 1.28 to 1.27. The ROUGE-1 score shows a minimal increase, rising from 24.9 to 25.0.

Table 4.1.3 - FLAN-T5-Base on BillSum, With 8000 Rows in Train Set and 1000 Rows in Test Set, After Training on XSum

<i>Epoch</i>	<i>Training Loss</i>	<i>Validation Loss</i>	<i>Rouge1</i>
1	1.2891	1.195633	24.9941
2	1.2783	1.184432	25.0385
3	1.2178	1.177504	25.0069
4	1.2701	1.174383	25.009

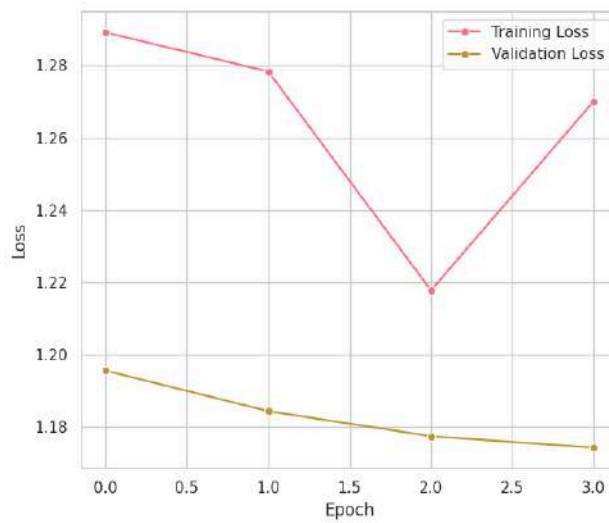


Fig. 4.1.5 - Loss graph for Table 4.1.3

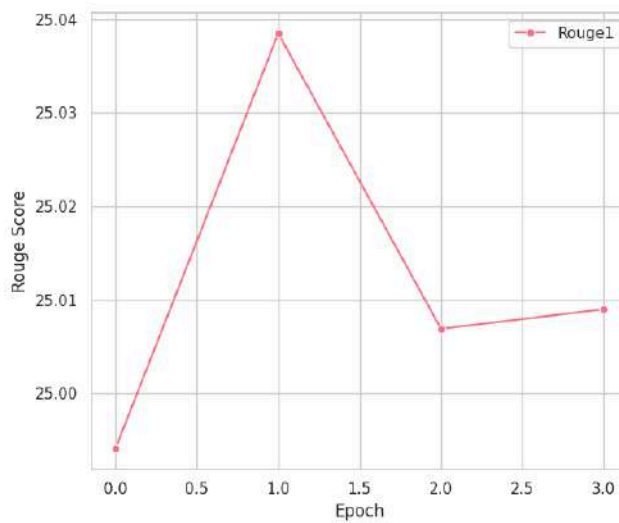


Fig. 4.1.6 - Rouge Score graph for Table 4.1.3

In Table 4.1.4, FLAN-T5 exhibits performance similar to that observed in Table II. While the validation loss falls from 1.8 to 1.7, the training loss decreases somewhat from 2.0 to 1.9. The ROUGE-1 score shows a modest increase from 17.9 to 18.1.

Table 4.1.4 - Flan-T5-Base on BillSum With 18.9K Rows in Training and 3.27K Rows in Test Set.

<i>Epoch</i>	<i>Training Loss</i>	<i>Validation Loss</i>	<i>Rouge1</i>
1	2.0935	1.870951	17.9862
2	1.9908	1.823429	18.1394
3	2.0161	1.802811	18.1797
4	1.9716	1.797551	18.1915

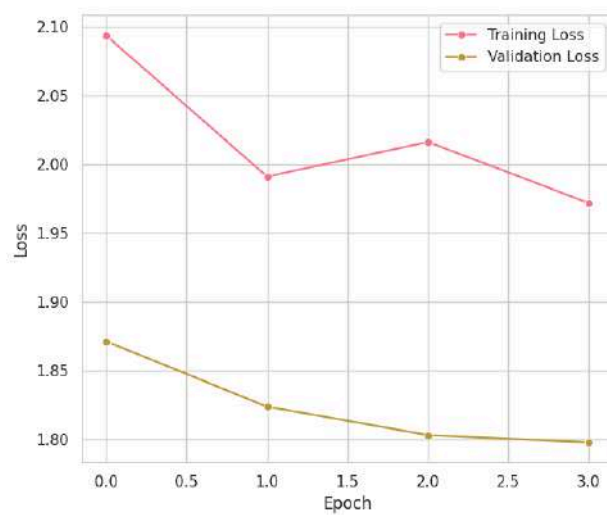


Fig. 4.1.7 - Loss graph for Table 4.1.4

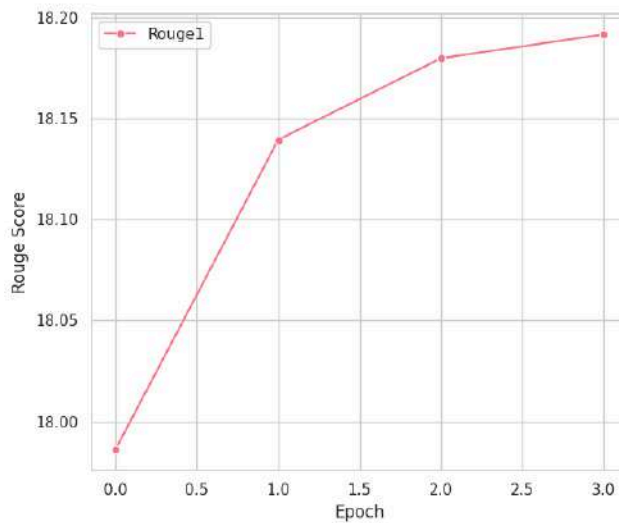


Fig. 4.1.8 - Rouge Score for Table 4.1.4

In Table 4.1.5, FLAN-T5 performs comparably to BART-Large as shown in Table I. The validation loss falls from 1.9 to 1.8, while the training loss falls from 2.1 to 2.0. The ROUGE-1 score significantly improves, rising from 39.1 to 40.8.

Table 4.1.5 - FLAN-T5-Base on BillSum After Hyperparameter Tuning

<i>Epoch</i>	<i>Training Loss</i>	<i>Validation Loss</i>	<i>Rouge1</i>
1	2.1837	1.926986	39.1855
2	2.126	1.88247	39.6472
3	2.0353	1.863689	41.1076
4	2.0528	1.857668	40.8503

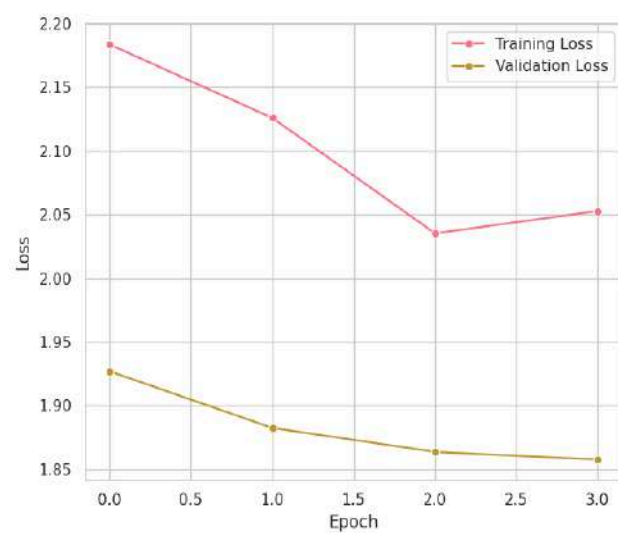


Fig. 4.1.9 - Loss graph for Table 4.1.5

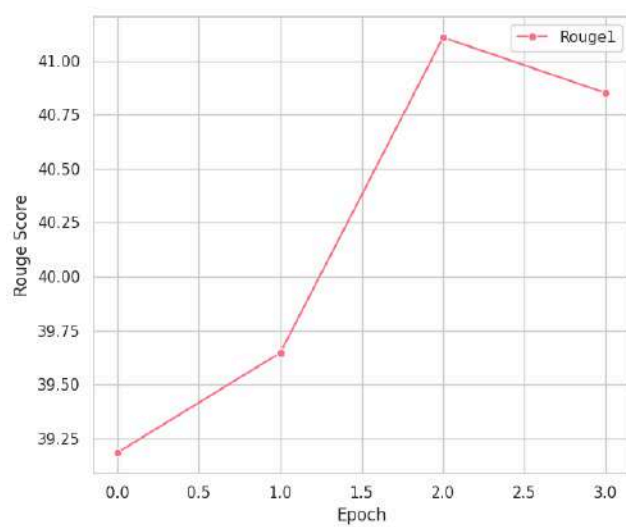


Fig. 4.1.10 - Rouge Score for Table 4.1.5

CHAPTER 5

CONCLUSION, FUTURE SCOPE, SOCIAL IMPACT

5.1 Conclusion

Through the integration of bibliometric analysis and practical testing on transformer-based models, this thesis offers a dual-perspective investigation of the developing subject of text summarization. The two-pronged approach has provided valuable insights into both the macro-level research landscape and the micro-level technical performance improvements in domain-specific summarisation tasks.

From a bibliometric standpoint, the field of text summarisation has witnessed significant growth over the past two decades, with a surge in research output particularly after 2015. High-impact papers, eminent institutions, and significant contributors were identified through the examination of 983 publications from the Web of Science database. Journals like *Expert Systems with Applications* and *IEEE Access* have been important publication sites, and nations including the US, China, and India have been significant players in this field. There has been a noticeable shift from extractive techniques like TextRank to sophisticated abstractive strategies that make use of transformer structures like BERT and GPT. Also, modern patterns that demonstrate a develop and differentiated consider space incorporate topic moves toward multimodal, multilingual, and real-time summarization.

In terms of experimentation, this thesis accesses how well two transformer models—FLAN-T5 Base and BART-Large—summarize data from the BillSum dataset. Agreeing to preparatory comes about, FLAN-T5 performed discernibly more awful than BART-Large. Negligible picks up were gotten when FLAN-T5's execution was endeavored to be moved forward by extending the dataset and doing more pretraining on the XSum dataset. Targeted hyperparameter optimization, namely beam width and generation length modifications, resulted in a significant improvement in performance, with ROUGE-1 scores increasing from 18.3 to 40.85, which essentially matched BART-Large. These discoveries cast question on the idea that more current models naturally perform way better than more seasoned ones without satisfactory tuning and highlight the importance of model-specific arrangement.

This proposition basically emphasizes how academic progression and real-world optimization are associated within the field of content summarization. It

affirms that both comprehending inquire about patterns and moving forward the fundamental advances can lead to noteworthy breakthroughs. For up and coming researchers and experts trusting to improve in this dynamic and critical field, the combined discoveries give a street outline.

5.2 Future Scope

There are several directions for future research in the developing field of text summarization, especially in improving model flexibility across various domains with little supervision. There are several directions for future research in the developing field of text summarization, especially in improving model flexibility across various domains with little supervision. Furthermore, utilizing multimodal data sources and cross-lingual transfer learning offers the possibility of developing more resilient and broadly applicable summarization systems, opening the door for context-aware and adaptive language models.

5.3 Social Impact

Better text summarizing models could have a significant impact on how people and organizations use information, encouraging accessibility and well-informed choices. Reliable summary can improve transparency, lessen information overload, and facilitate quicker understanding of important papers in the legal, medical, and governmental sectors. Such innovations promote inclusive technology growth and more equitable access to information by reducing the digital divide and enabling the spread of knowledge in under-resourced languages or geographical areas.

REFERENCES

- [1] ROUGE: A Package for Automatic Evaluation of Summaries. (Lin, 2004)
- [2] Chung, Hyung Won, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li et al. "Scaling instruction-finetuned language models." *Journal of Machine Learning Research*, 25(70), 2024, pp. 1–53.
- [3] Raffel, Colin, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. "Exploring the limits of transfer learning with a unified text-to-text transformer." *Journal of Machine Learning Research*, 21(140), 2020, pp. 1–67.
- [4] Zhang, Shengyu, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li et al. "Instruction tuning for large language models: A survey." *arXiv preprint arXiv:2308.10792* (2023).
- [5] Lewis, Mike, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension." *Annual Meeting of the Association for Computational Linguistics*, 2019.
- [6] Liu, Xueqing, and Chi Wang. "An empirical study on hyperparameter optimization for fine-tuning pre-trained language models." *arXiv preprint arXiv:2106.09204* (2021).
- [7] Ülker, M., and Özer, A. B. "The Bart-based Model for Scientific Articles Summarization." *Journal of Universal Computer Science*, 30(13), 2024, pp. 1807–1828. <https://doi.org/10.3897/jucs.115121>
- [8] Abdul Ghafoor Etemad, Ali Imam Abidi, and Megha Chhabra. "Fine-Tuned T5 for Abstractive Summarization." *International Journal of Performability Engineering*, 17(10), 2021, pp. 900–906.
- [9] Waskom, M. L. "Seaborn: Statistical data visualization." *Journal of Open Source Software*, 6(60), 2021, Article 3021. <https://doi.org/10.21105/joss.03021>
- [10] Kornilova, Anastassia, and Vladimir Eidelman. "BillSum: A Corpus for Automatic Summarization of US Legislation." In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, 2019, pp. 48–56. Hong Kong, China: Association for Computational Linguistics.

- [11] Zhang, Tianyi, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. "BERTScore: Evaluating text generation with BERT." *arXiv preprint arXiv:1904.09675* (2019).
- [12] Hu, Edward J., Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. "LoRA: Low-Rank Adaptation of Large Language Models." *International Conference on Learning Representations (ICLR)*, 1(2), 2022, Article 3.
- [13] Wolf, Thomas, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac et al. "HuggingFace's Transformers: State-of-the-Art Natural Language Processing." *arXiv preprint arXiv:1910.03771* (2019).
- [14] Narayan, Shashi, Shay B. Cohen, and Mirella Lapata. "Don't Give Me the Details, Just the Summary! Topic-Aware Convolutional Neural Networks for Extreme Summarization." *arXiv preprint arXiv:1808.08745* (2018).
- [15] Kaushal, Anukriti, Chia-Chen Lin, Rishabh Chauhan, and Rajeev Kumar. "Charting the Growth of Text Summarisation: A Data-Driven Exploration of Research Trends and Technological Advancements." *Applied Sciences*, 14(23), 2024, Article 11462. <https://doi.org/10.3390/app142311462>
- [16] El-Kassas, W.S.; Salama, C.R.; Rafea, A.A.; Mohamed, H.K. Automatic text summarization: A comprehensive survey. *Expert Syst. Appl.* **2021**, *165*, 113679.
- [17] Gambhir, M.; Gupta, V. Recent automatic text summarization techniques: A survey. *Artif. Intell. Rev.* 2017, *47*, 1–66.
- [18] Mihalcea, R.; Tarau, P. Textrank: Bringing order into text. In Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, Barcelona, Spain, 25–26 July 2004; pp. 404–411.
- [19] Erkan, G.; Radev, D. LexRank: Graph-based lexical centrality as salience in text summarization. *J. Artif. Intell. Res.* 2004, *22*, 457–479.



DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Shahbad Daultpur, Main Bawana Road, Delhi-42

PLAGIARISM VERIFICATION

Title of the Thesis _____

Total Pages _____ Name of the Scholar _____

Supervisor (s)

(1) _____

(2) _____

(3) _____

Department _____

This is to report that the above thesis was scanned for similarity detection. Process and outcome is given below:

Software used: _____ Similarity Index: _____, Total Word Count: _____

Date: _____

Candidate's Signature

Signature of Supervisor(s)

0% detected as AI

The percentage indicates the combined amount of likely AI-generated text as well as likely AI-generated text that was also likely AI-paraphrased.

Caution: Review required.

It is essential to understand the limitations of AI detection before making decisions about a student's work. We encourage you to learn more about Turnitin's AI detection capabilities before using the tool.

Detection Groups



0 AI-generated only 0%

Likely AI-generated text from a large-language model.



0 AI-generated text that was AI-paraphrased 0%

Likely AI-generated text that was likely revised using an AI-paraphrase tool or word spinner.

Disclaimer

Our AI writing assessment is designed to help educators identify text that might be prepared by a generative AI tool. Our AI writing assessment may not always be accurate (it may misidentify writing that is likely AI generated as AI generated and AI paraphrased or likely AI generated and AI paraphrased writing as only AI generated) so it should not be used as the sole basis for adverse actions against a student. It takes further scrutiny and human judgment in conjunction with an organization's application of its specific academic policies to determine whether any academic misconduct has occurred.

Frequently Asked Questions

How should I interpret Turnitin's AI writing percentage and false positives?

The percentage shown in the AI writing report is the amount of qualifying text within the submission that Turnitin's AI writing detection model determines was either likely AI-generated text from a large-language model or likely AI-generated text that was likely revised using an AI-paraphrase tool or word spinner.

False positives (incorrectly flagging human-written text as AI-generated) are a possibility in AI models.

AI detection scores under 20%, which we do not surface in new reports, have a higher likelihood of false positives. To reduce the likelihood of misinterpretation, no score or highlights are attributed and are indicated with an asterisk in the report (*%).

The AI writing percentage should not be the sole basis to determine whether misconduct has occurred. The reviewer/instructor should use the percentage as a means to start a formative conversation with their student and/or use it to examine the submitted assignment in accordance with their school's policies.

What does 'qualifying text' mean?

Our model only processes qualifying text in the form of long-form writing. Long-form writing means individual sentences contained in paragraphs that make up a longer piece of written work, such as an essay, a dissertation, or an article, etc. Qualifying text that has been determined to be likely AI-generated will be highlighted in cyan in the submission, and likely AI-generated and then likely AI-paraphrased will be highlighted purple.

Non-qualifying text, such as bullet points, annotated bibliographies, etc., will not be processed and can create disparity between the submission highlights and the percentage shown.



Rishabh

MtechThesisContent_merged-4-43.pdf

 Delhi Technological University

Document Details

Submission ID

trn:oid:::27535:97990414

Submission Date

May 27, 2025, 5:09 PM GMT+5:30

Download Date

May 27, 2025, 5:10 PM GMT+5:30

File Name

MtechThesisContent_merged-4-43.pdf

File Size

9.7 MB

40 Pages

5,522 Words

31,663 Characters

13% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

Filtered from the Report

- Bibliography
- Quoted Text
- Cited Text

Match Groups

-  **56 Not Cited or Quoted 13%**
Matches with neither in-text citation nor quotation marks
-  **0 Missing Quotations 0%**
Matches that are still very similar to source material
-  **0 Missing Citation 0%**
Matches that have quotation marks, but no in-text citation
-  **0 Cited and Quoted 0%**
Matches with in-text citation present, but no quotation marks

Top Sources

- 8%  Internet sources
- 6%  Publications
- 7%  Submitted works (Student Papers)

Integrity Flags

0 Integrity Flags for Review

No suspicious text manipulations found.

Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.

Match Groups

- 56 Not Cited or Quoted 13%**
Matches with neither in-text citation nor quotation marks
- 0 Missing Quotations 0%**
Matches that are still very similar to source material
- 0 Missing Citation 0%**
Matches that have quotation marks, but no in-text citation
- 0 Cited and Quoted 0%**
Matches with in-text citation present, but no quotation marks

Top Sources

- 8% Internet sources
- 6% Publications
- 7% Submitted works (Student Papers)

Top Sources

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

1	Internet	www.mdpi.com	2%
2	Internet	123dok.net	1%
3	Submitted works	University of Wollongong on 2021-06-24	1%
4	Publication	Anukriti Kaushal, Chia-Chen Lin, Rishabh Chauhan, Rajeev Kumar. "Charting the ...	<1%
5	Submitted works	Gisma University of Applied Sciences GmbH on 2025-03-25	<1%
6	Internet	etd.lib.metu.edu.tr	<1%
7	Internet	www.collectionscanada.ca	<1%
8	Submitted works	Liverpool John Moores University on 2024-04-10	<1%
9	Internet	acikbilim.yok.gov.tr	<1%
10	Submitted works	University of Wollongong on 2023-12-08	<1%

11	Internet	project.obiaxs.com	<1%
12	Internet	text-id.123dok.com	<1%
13	Internet	www.spiedigitallibrary.org	<1%
14	Submitted works	Aston University on 2018-04-27	<1%
15	Publication	Odeyemi, Glory. "Finetuning Encoder-Decoder Transformer Models for Text Sum..."	<1%
16	Internet	kipdf.com	<1%
17	Internet	export.arxiv.org	<1%
18	Internet	www.academicjournals.org	<1%
19	Submitted works	University of Wolverhampton on 2024-12-19	<1%
20	Internet	www.ir.juit.ac.in:8080	<1%
21	Publication	Lucke, Michael. "Using Text-Based Representations of Knowledge Graphs to Impr..."	<1%
22	Submitted works	The University of Manchester on 2020-09-01	<1%
23	Submitted works	University of Westminster on 2022-07-25	<1%
24	Submitted works	Swiss School of Business and Management - SSBM on 2024-05-16	<1%

25	Submitted works	University of Sheffield on 2023-09-13	<1%
26	Internet	aclanthology.org	<1%
27	Internet	open.library.ubc.ca	<1%
28	Internet	www.docstoc.com	<1%
29	Submitted works	The Robert Gordon University on 2024-12-16	<1%
30	Publication	Ingrid N. Pinto-López, Cynthia M. Montaudon-Tomas. "chapter 3 Distance Learnin...	<1%
31	Publication	Zmandar, Nadhem. "Multilingual Financial Text Summarisation", Lancaster Unive...	<1%

LIST OF PUBLICATIONS

Kaushal, Anukriti, Chia-Chen Lin, Rishabh Chauhan, and Rajeev Kumar. 2024. "Charting the Growth of Text Summarisation: A Data-Driven Exploration of Research Trends and Technological Advancements" *Applied Sciences* 14, no. 23: 11462. <https://doi.org/10.3390/app142311462>



Kaushal, Anukriti, and Rishabh Chauhan. "Hyperparameter Tuning for Domain-Specific Text Generation: A Case Study with FLAN-T5." Delhi Technological University, 2025.



 3rd International Conference on Inventive Computing and Informatics ICICI 2025 4-6, June 2025 - https://icinci.com/2025/ - iciciconfcontact@gmail.com	
<i>Payment Receipt</i>	
Receipt No.	ICICI-2025-234
Paper Title	Hyperparameter Tuning for Domain-Specific Text Generation: A Case Study with FLAN-T5
Paid by	Anukriti Kaushal, Rishabh Chauhan
Amount Paid (in Words)	Rupees Eleven Thousand Five Hundred only
Amount Paid	₹ 11,500/-
  Dr. Usha Desai Conference Chair	

 **Transaction Successful**
03:09 PM on 22 May 2025

Paid to

 **CONFYY** ₹11,500
ConferenceFee10.rzp@icici

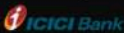
 **Transfer Details** ^

Message
PaymenttoConferenceFee

Transaction ID
T2505221509395737585867

Debited from

 **XXXXXX9705** ₹11,500
UTR: 106634774334

Powered by
 



Article

Charting the Growth of Text Summarisation: A Data-Driven Exploration of Research Trends and Technological Advancements

Anukriti Kaushal, Chia-Chen Lin, Rishabh Chauhan and Rajeev Kumar





Article

Charting the Growth of Text Summarisation: A Data-Driven Exploration of Research Trends and Technological Advancements

Anukriti Kaushal ^{1,*}, Chia-Chen Lin ^{2,*}, Rishabh Chauhan ¹ and Rajeev Kumar ¹

¹ Department of Computer Science and Engineering, Delhi Technological University, Delhi 110042, India; dr.smolderrishabhchauhan@gmail.com (R.C.); rajeevkumar@dtu.ac.in (R.K.)

² Department of Computer Science and Information Engineering, National Chin-Yi University of Technology, Taichung 411030, Taiwan

* Correspondence: anukritikaushal30@gmail.com (A.K.); ally.cclin@ncut.edu.tw (C.-C.L.)

Abstract: Text summarisation plays a pivotal role in efficiently processing large volumes of textual data, making it an indispensable tool across diverse domains such as healthcare, legal, education, and journalism. It addresses the challenge of information overload by condensing or generating concise, meaningful summaries that improve decision-making, enhance accessibility, and save valuable time. Advances in artificial intelligence continue to propel the growth of text summarisation research, particularly with the evolution from traditional extractive approaches to cutting-edge abstractive models like BERT and GPT, as well as emerging innovations in multimodal and multilingual summarisation. To trace the development of this field, this study integrates bibliometric analysis and an in-depth survey, leveraging data from the Web of Science database to explore citation trends, uncover influential contributors, and highlight emerging research areas. Furthermore, bibliometric and critical evaluations are employed to outline strategic pathways and propose future directions for the continued advancement of the field. By incorporating sophisticated visualisation tools such as VOSviewer and RawGraphs, the analysis provides an enriched understanding of the field's trajectory, identifying significant methodologies, landmark contributions, and existing gaps. This comprehensive exploration not only underscores the progress achieved in text summarisation but also serves as an invaluable resource for shaping forthcoming research endeavours and inspiring innovation in this dynamic area of study.

Keywords: bibliometric analysis; WoS; text summarisation; NLP; LLM



Citation: Kaushal, A.; Lin, C.-C.; Chauhan, R.; Kumar, R. Charting the Growth of Text Summarisation: A Data-Driven Exploration of Research Trends and Technological Advancements. *Appl. Sci.* **2024**, *14*, 11462. <https://doi.org/10.3390/app142311462>

Academic Editor: Andrea Prati

Received: 3 November 2024

Revised: 4 December 2024

Accepted: 5 December 2024

Published: 9 December 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

For several decades, text summarisation has attracted a lot of attention in the natural language processing (NLP) community. Text summarisation focuses on condensing large documents into shorter versions that retain the key points, either through extracting important sentences or generating concise summaries. Large-scale text summarisation is an essential skill for content management, information retrieval, and decision-making in a variety of fields. Also, NLP techniques, including text summarisation, have been extensively used in analysing online customer reviews to extract valuable insights and improve decision-making processes [1]. Text summarisation techniques have undergone significant evolution over time, driven by advancements in computational power, availability of data, and NLP [2] methodologies. The primary approaches include extractive, abstractive, hybrid, query-based, multi-document, and domain-specific summarisation.

The early 1990s approaches to text summarisation were extractive, using statistical techniques [3] requiring redundancy removal techniques in the post-processing phase to improve summary quality. This involved selecting key sentences or phrases directly from the original text, initially relying on statistical methods such as frequency-based scoring and TF-IDF (Term frequency-inverse document frequency) [4]. Without abstraction, they select statements without rephrasing, resulting in redundant or disconnected

summaries with poor contextual flow. Semantic comprehension, domain-specific reasoning, and implicit meanings are difficult for these approaches to handle in complicated or lengthy texts [5]. Over time, graph-based methods like TextRank [6] and machine learning models emerged [7], culminating in transformer-based models such as BERTSUM [8] in recent years.

Abstractive summarisation [9], on the other hand, generates new sentences by rephrasing content, requiring a deeper understanding of the text. Early rule-based systems evolved into sequence-to-sequence models with attention mechanisms [10,11], and with the introduction of transformer architectures in 2018 they have emerged as the most significant and recent advancement in text summarisation. Transformers have revolutionised the field, e.g., BERT (Bidirectional Encoder Representations from Transformers.) [8] in 2020 and GPT (Generative Pre-trained Transformer) [12] from 2018 to 2023 and PEGASUS (Pre-training with extracted gap sentences for abstractive summarisation) [13] in 2020, offering greater coherence and faithfulness. Text summarisation is a rapidly evolving topic with a growing body of research; therefore, it is imperative to perform a thorough bibliometric study to identify the trends, patterns, and key players influencing this field.

1.1. Survey of Surveys

This study builds upon and extends previous surveys on text summarisation by incorporating a comprehensive bibliometric analysis, examining research trends, collaborations, and thematic evolution in the field. Unlike prior works, such as Gambhir and Gupta [14] (2017) and El-Kassas et al. [15] (2021), which primarily focused on summarisation techniques and advancements in deep learning models, this paper integrates quantitative methods using tools like VOSviewer (V_1.6.18) and RawGraphs (V_2.0) to visualise co-authorship networks, citation patterns, and keyword co-occurrences. By analysing 983 publications, including several recent and influential papers, this study provides new insights into emerging areas such as domain-specific summarisation, real-time applications, and ethical considerations, which were not extensively covered in earlier reviews. Additionally, it discusses underexplored topics like multimodal [16] and multilingual [17] summarisation, dialogue summarisation, and explainability as a structured framework for future research. This integrative approach makes this survey unique, bridging the gap between traditional reviews of summarisation techniques and a broader analysis of the research landscape, while incorporating the latest advancements [18,19] and trends in the field. This survey of surveys in Table 1 aims to provide a comprehensive overview of significant studies that have examined various text summarisation techniques or pertinent to text summarisation methodologies. By analysing existing surveys, we can identify the focus areas, methodologies, and findings presented in each, providing insight into the field's progression and future research directions. To further categorise these reviews, we present a classification system based on various parameters that designate research areas as 'less explored areas' (L), 'moderately explored areas' (M), and 'highly explored areas' (H). For understanding where research efforts have been concentrated and where opportunities for further exploration lie, this classification is helpful.

Table 1. Summary of relevant literature based on various scopes and references.

Scope/References	[20]	[21]	[22]	[23]	[24]	[25]	This Survey
Taxonomy/Overview	M	M	H	H	H	H	H
Publication Connections	L	L	H	M	M	H	H
Topics	M	L	M	M	H	M	H
Keyword Analysis	L	L	L	M	M	H	H
Techniques	H	M	M	M	L	H	H
Key Innovators and Scholarly Networks	M	M	L	M	H	H	H
Citation Insights	L	M	L	L	M	M	H
Future Directions	M	M	M	H	H	M	H

1.2. Research Aim and Questions

Bibliometric analysis is a useful tool for understanding research trends, assessing influence, and visualising collaborations. It helps in finding prominent topics, leading researchers, and impactful publications, offering valuable insights into the development of a field. Researchers leverage it to analyse productivity, choose reputable journals, and identify collaboration opportunities. Here are some research questions we have proposed to better understand our study:

RQ1: How has research evolved in text summarisation over the past two decades?

RQ2: Which authors, institutes, and countries are collaborating and contributing most in this field?

RQ3: Which papers are deemed most important by researchers for exploring this field?

RQ4: In which direction should researchers keep an eye for further research?

The growth of publications over the years is analyzed in RQ1 through a bar graph, showcasing the increasing number of publications. This visualisation helps assess the influence of text summarisation on researchers and its potential for future growth. Additionally, a chord diagram in Section 5 highlights which publishers were more active in specific years. Understanding the interests of different publishers in particular fields enables researchers to identify reputable outlets, track research trends, and refine their publication strategies.

To address RQ2, network graphs are plotted to visualize collaboration among countries, authors, and institutions as detailed in Sections 3 and 4. These graphs provide insights into the geographical and institutional impact of the field, measuring relationships between entities using attributes like Total Link Strength. This helps highlight the level of collaboration and influence across regions and organisations.

For RQ3, we identify influential studies and research papers, as discussed in the Survey Findings section. This analysis helps new researchers familiarize themselves with prominent work in the field, building a strong foundation. Finally, RQ4 explores how researchers can align their interests with major contributors or focus on under-represented regions to raise awareness and foster balanced growth within the domain.

1.3. Key Contribution

This study offers a thorough bibliometric analysis of the text summarisation research domain, emphasising significant trends, prominent authors, and novel approaches. The following are the main contributions of this study:

- This work involves the collection, analysis, and representation of a comprehensive dataset of publications related to text summarisation research using an efficient bibliometric technique.
- This study examines contemporary review articles to pinpoint shortcomings in prior surveys and address their limitations, providing a meta-analytical viewpoint on research trends.
- It offers a comprehensive analysis of key papers, presenting a thorough overview of significant contributions, methodologies, key features, and Limitations of research.
- The study performs a thorough evaluation of extensively cited and contemporary publications, evaluating their impact and contribution to the advancement of research.
- This study offers practical insights derived from the results of the bibliometric analysis. These insights reveal emerging trends, unexplored areas, and critical issues for future investigation.
- This study serves as an essential resource for scholars, supporting them in traversing the dynamic domain of text summarisation research through the integration of bibliometric techniques with comprehension and critical evaluation.

The remainder of this paper is organised as follows: Section 2 details the methodology, including the process and motivation for collecting data on text summarisation research papers and performing a bibliometric analysis. Section 3 presents an analysis of publications, covering trends in growth, types, WoS categories, titles, locations, institutions,

and authors. Section 4 offers a citation analysis based on contributions from countries, organisations, researchers, and articles. Section 5 examines international cooperation and publisher collaborations over the years. Section 6 outlines the survey findings, discusses future directions in text summarisation and addresses the limitations of the study. Section 7 concludes the paper.

2. Methodology

A quantitative technique for assessing and analysing academic literature based on bibliographic data is called bibliometric analysis [26,27]. In this study, journals, articles, authors, and keywords are among the publications that are examined for patterns, trends, and correlations. Through the analysis of citation trends, scholars can pinpoint significant writers and establishments, promoting cooperation and comprehension of subject matter competence. Additionally, bibliometric indicators, such as journal impact factors and citation counts, enable the assessment of the quality and influence of research, helping scholars and decision-makers determine the importance of particular studies or publications. Mapping collaborative networks between researchers and institutions is made easier with the use of bibliometric analysis, which is essential for encouraging collaboration and knowledge exchange. Furthermore, researchers can identify study gaps and focus on understudied aspects of text summarisation by employing bibliometric analysis to identify areas with fewer publications or citations.

Data Collection

The widely used WoS provides access to scientific information, impact assessments, and collections going back to the Start Year. The WoS Core Collection stands out among the other WoS resources because it provides researchers with a vast collection of credible journals and publications in addition to relevant data that can be easily incorporated into bibliometric analysis tools. Four important databases are included in this collection: A&HCI (Arts and Humanities Citation Index), CPCI-S (Proceedings Citation Index-Science), SCI-EXPANDED (Science Citation Index Expanded), and SSCI (Social Sciences Citation Index).

We have examined each of the aforementioned databases to guarantee the precision and completeness of our data. SCI-EXPANDED, SSCI, CPCI-S, and A&HCI make up the most accurate collection of databases according to our analysis. A PRISMA diagram Figure 1 shows the search strategy used on the WoS database and other sources, including and excluding results. WoS maintains popular repositories and indexing services for multiple disciplines. Repositories include IEEE Xplore, ACM Digital Library, Springer-Link, and others. It includes Elsevier's Scopus, Nature Publishing Group, and Wiley Online Library journals and conferences. WoS handles duplicates with robust metadata standardisation. WoS assigns a unique identifier (UT) to each publication, ensuring that the same record never counts twice and ensuring bibliometric analyses are accurate and free of redundancy. The search query given as "TS = ("Text summarization" OR "text summarizes" OR "text summarizing" OR "Text summarisation" OR "text summarises" OR "text summarising")" is used TS (Topic Search) to collect all possible papers in the title, abstract, author keywords, and keywords plus. Furthermore, the query includes "text summarisation", "text summarises", "text summarising", "text summarises", and many other terms separated by the OR operator. The query captured all possible spellings and grammatical forms of the term "text summarisation", including both common and uncommon variations. After a thorough search, the query returns the best filter results. The use of quotation marks ensured that each term matched exactly to the phrase.

The researchers have published 983 articles from 2004 to September 2024, demonstrating consistent academic productivity centred on recent advancements and trends in text summarisation and related disciplines. These works have garnered 18,392 citations, with 14,833 excluding self-citations, highlighting their esteemed reputation and considerable influence within the academic community. The publications, with an average of 18.37 citations per article, demonstrate the significance of their contributions, which span

338 journals and involve collaborations with 2952 authors. The researcher's influence is demonstrated by an impressive H-index of 60, with at least 60 publications cited 60 times or more. Furthermore, their work has been referenced in 12,285 articles, with 11,657 excluding self-citations, indicating a significant and enduring influence in the domains of computer science, artificial intelligence, and associated fields. The research is exclusively presented in English, emphasising original and critical writings, with document types further classified as filters to improve scholarly engagement. This body of work shows how important the researcher is to advancing knowledge and driving innovation in their field.

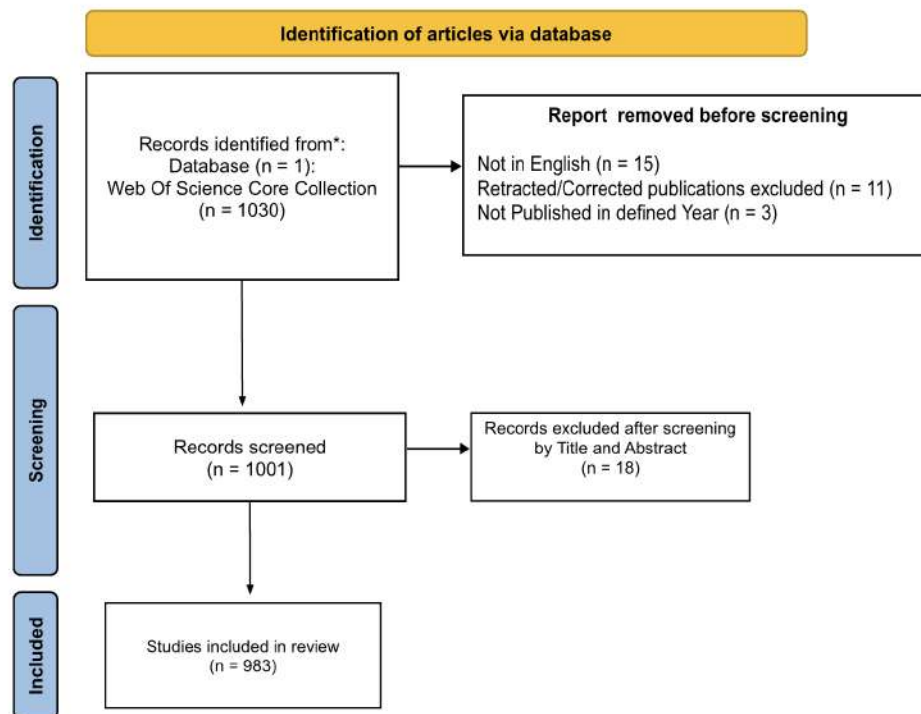


Figure 1. Prisma diagram for search query (here ‘*’ represents the number of records identified from each database searched).

3. Analysis of Publication Structure

We carried out a thorough analysis of research data covering several aspects, such as annual publication patterns, publication categories, prospective future research directions, authors, and countries of origin. We also looked at the collaboration network component with authors, nations, and organisations as data points.

The paper titled “The Automatic Creation of Literature Abstracts” [10] by H.P. Luhn in 1958 focused on using word frequency and sentence structure to summarise texts. This paper serves as the foundation for future developments in the field of text summarisation. The following points are taken into account in this analysis:

1. The quantity of documents released between March 2004 and September 2024.
2. Document types such as article and review article.
3. Most prestigious publications sources.
4. Artificial Intelligence, Electrical Engineering, Computer Science, and other fields are among the top research areas.
5. The leading nations and areas.
6. Graphs showing the top nations and the quantity of documents generated annually.
7. Prominent researchers, their nations, and the documents they have created are mapped out.

VOSviewer is used to examine and explore maps created from network data. VOSviewer is also used to investigate different types of cooperative relationships. Among these are collaborations between

1. Nations and areas with the most influence
2. The leading organisation
3. The most significant journal.

Finally, we present information about publishers over the years using a chord diagram created with Raw Graphs, an online graph-making tool.

A comprehensive analysis of the publication environment is given in this section, which takes into consideration several characteristics such as annual publication patterns, high publication rates in countries/regions, prolific institutions, and well-known authors.

3.1. An Analysis of Publications on an Annual Basis

The WoS database enables the analysis of current and emerging research domains covered in published studies. As shown in Figure 2, a topic-specific search using WoS data yielded 983 papers from the past 20 years, including the current year. In the figure, the y-axis represents the number of publications, while the x-axis shows the corresponding years. The data collection spans from 2015 to the present. The diagram reveals a steady increase in publications over time, indicating that the research domain is experiencing significant growth and is likely to continue expanding in the future.

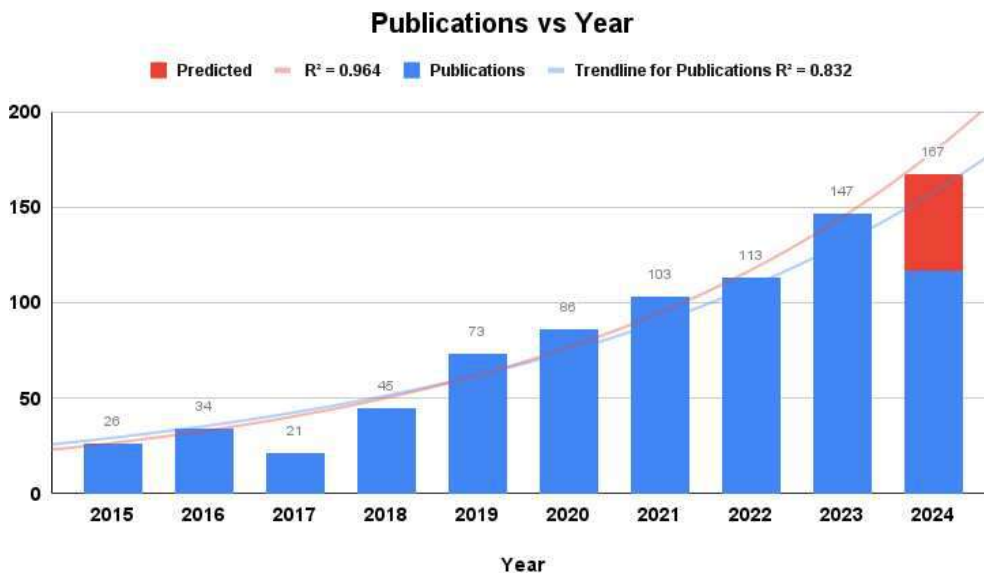


Figure 2. Annual publication growth graph from 2015 to 2024, showing the annual increase in publications.

3.2. Publications Examined: Type of Document

Figure 3 shows an analysis of the WoS from the 983 total documents acquired to analyse the type of publications. As can be seen, articles make up the majority of publishing types, accounting for 941 and 95.72% of all publications. In addition, there are review articles 41 (4.171%), early access 25 (2.543%), proceeding papers 40 (4.069%), and letters, among others.

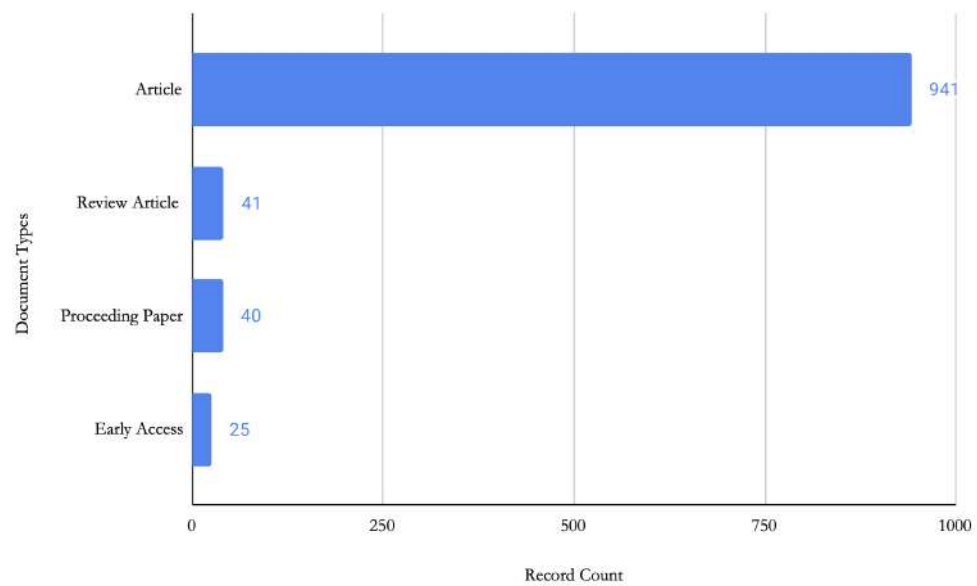


Figure 3. The publications' top ten categories.

3.3. Examining Publications: WoS Category Analysis

The publication organisation according to WoS category branch field is covered in this subsection.

Figure 4 illustrates the substantial influence of various domains on text summarisation. Among them, the Computer Science branch, particularly Artificial Intelligence, has the highest publication count (382), representing 38.861% of all publications. A closer analysis shows that the number of articles in Computer Science surpasses, and possibly overlaps, those in Electrical and Electronics Engineering by approximately 955. Within Computer Science, several subfields are represented, including Artificial Intelligence, Theory and Methods, Information Systems, Software Engineering, and Multidisciplinary Applications. Additionally, 34 papers are found in medical informatics, 80 in telecommunications, 46 in multidisciplinary engineering, 27 in linguistics, and others.

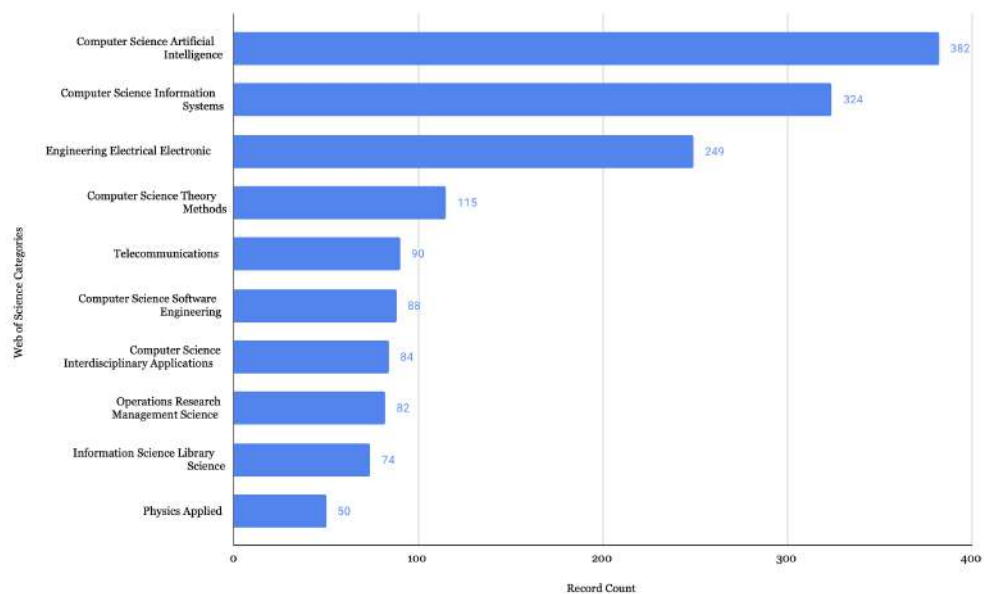


Figure 4. The primary focus areas denoted by the top 10 research directions in the publications.

3.4. Publications Examined: Titles of Publications

The top 10 published titles are displayed in Figure 5, from which it can be seen that EXPERT SYSTEMS WITH APPLICATIONS has the majority, with 68 articles, which is 6.918%. Second on the list is IEEE ACCESS 65 (6.612%). Subsequently, INFORMATION PROCESSING MANAGEMENT contains 40 articles, followed by APPLIED SCIENCES BASEL (31), MULTIMEDIA TOOLS AND APPLICATIONS (28), KNOWLEDGE BASED SYSTEMS (25), JOURNAL OF INTELLIGENT FUZZY SYSTEMS (22), LECTURE NOTES IN COMPUTER SCIENCE (16), and JOURNAL OF BIOMEDICAL INFORMATICS (15), among other source titles present.

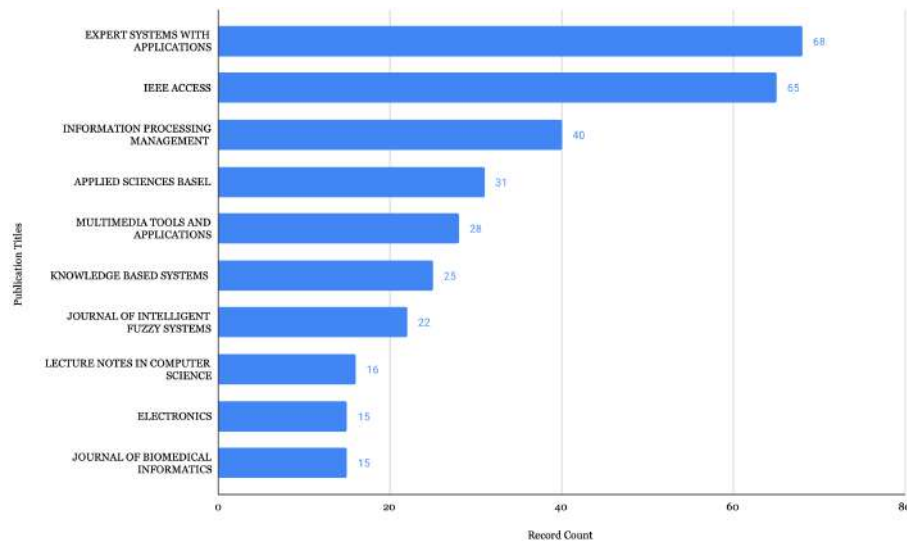


Figure 5. The publication titles of the primary sources.

3.5. Publications Examined: Well-Known Locations

The top countries and areas with the most publications are identified by the analysis. The distribution of publications across various places and nations is shown in Figure 6. In particular, the top-performing countries include South Korea (43 publications), Saudi Arabia (50 publications), Spain (61 publications), England (49 publications), the USA (129 publications), the People's Republic of China (215 publications), and India (172 publications).

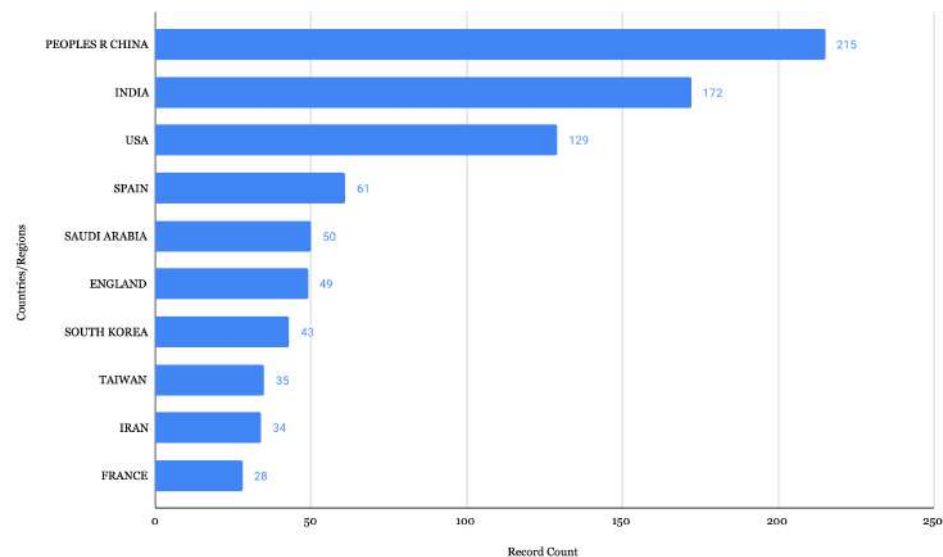


Figure 6. The top 10 productive places.

3.6. Examining Publications: Effective Institutions and Scholars

Based on a WoS study, the bar graph in Figure 7 shows the top ten organisations by affiliation type, illustrating contributions from a total of 1272 organisations. The top ten contributors are highlighted in Figure 8, with the “Indian Institute of Technology (IIT)” system being the most prolific, contributing 31 articles. It is followed by the “National Institute of Technology (NIT)” system, another Indian institution with 28 publications. Similarly, other notable institutions include the “Egyptian Knowledge Bank (EKB)” with 22 publications, the “Chinese Academy of Sciences” with 29, and “King Saud University” with 15 publications. The following contributions from the IIT system are particularly noteworthy and relevant to the field: “Text summarisation from Legal Documents: A Survey” [28]; “Extractive Single Document summarisation Using Multi-objective Optimization: Exploring Self-organized Differential Evolution, Grey Wolf Optimizer, and Water Cycle Algorithm” [29]; “MCRMR: Maximum Coverage and Relevancy with Minimal Redundancy Based Multi-document summarisation” [30] and “Automatic Scientific Document Clustering Using Self-organized Multi-objective Differential Evolution” [31].

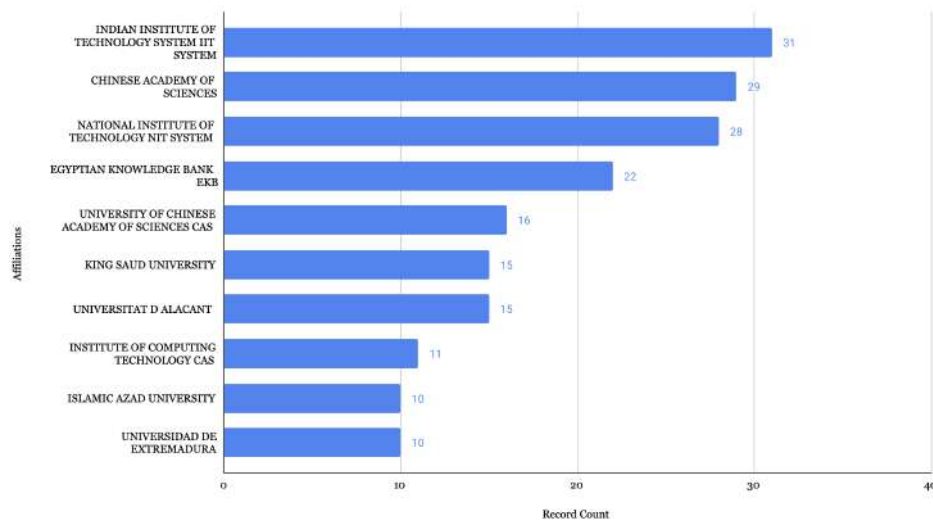


Figure 7. The organisations that contribute the most research articles.

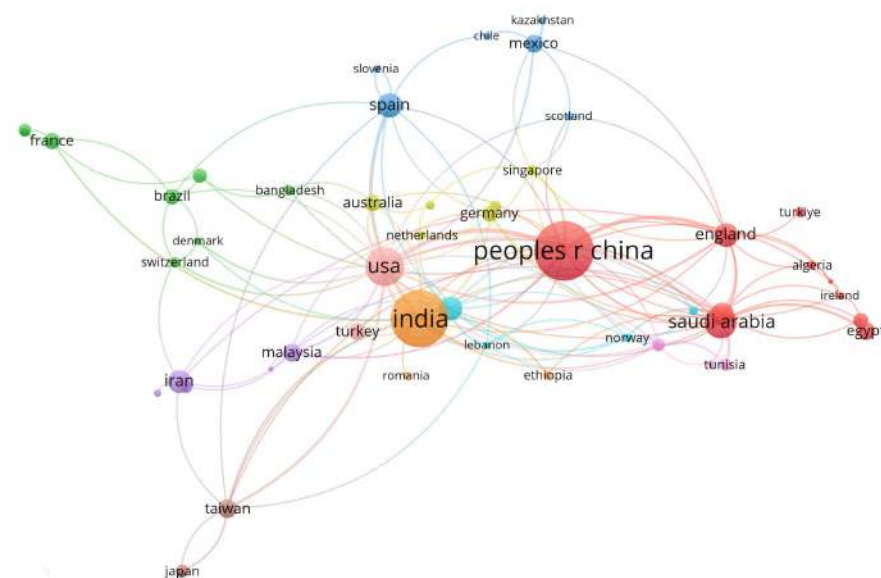


Figure 8. The collaborative network of places.

3.7. Analysing Publications: Authors and Institutions with Productivity

This subsection provides an in-depth analysis of the publishing environment, taking authorship and geographic affiliations into account. A good estimate of prolific authors using Lotka's Law can be calculated by the formula below:

$$\sqrt{\text{Number of authors}} = \sqrt{2952} = 54.3$$

Table 2 provides a detailed analysis of the top 20 authors with notably high productivity. Author Lloret E, who holds the highest publication count, collaborated with Palomar M on their most cited paper titled “Text summarisation in progress: a literature review” [32] with 154 citations. Moreover, the most cited work, “Extractive multi-document text summarisation using a multi-objective artificial bee colony optimisation approach” [33] with 62 citations, was coauthored by Sanchez-Gomez JM (9), Vega-Rodríguez MA (9), and Pérez CJ (9).

Table 2. A list of the writers who are most productive.

Rank	Authors	Record Count
1	Lloret E	14
2	Palomar M	10
3	Pérez CJ	9
4	Sanchez-gomez JM	9
5	Vega-rodríguez MA	9
6	Aliguliyev RM	8
7	Chatterjee N	8
8	García-hernández RA	8
9	Ledeneva Y	8
10	Gupta A	6
11	Idris N	6
12	Ko Y	6
13	Pakray P	6
14	Saha S	6
15	Seo J	6
16	Singh P	6
17	Zhang Y	6
18	Bhattacharyya P	5
19	Lins RD	5
20	Liu Y	5

4. Citation Structure Analysis

The purpose of this study is to illustrate how cooperatively different places, institutions, and scholars are working along this vertical axis. A VOSviewer was used to make this research easier to conduct. To get a thorough grasp, let us now examine each component in more detail.

Within the field of text summarisation research, bibliometric analysis with tools such as VOSviewer involves a number of key terms that need to be understood in order to interpret research findings, collaboration patterns, and the scholarly impact of the field's dynamic and ever-evolving field. One measure used to estimate research output is the Key Term Publication Count (P). Citation (C) indicates how many citations a specific author, document, or keyword has gotten. It provides a normalised metric that aids in evaluating the average influence of academic publications in relation to their individual publication counts. The reference link indicates the number of co-citations, and the total link strength, or ‘TLS’, quantifies the frequency of citations to the co-cited article. The degree or strength of the linkages between two entities, such as nations or institutions, is described by Link Strength (LS) in VOSviewer. On the other hand, TLS considers an entity's complete network of relationships and the total strength of all connections connected to it.

Each colour in the VOSviewer visualisation denotes a unique group or community of objects that show a lot of likeness. Depending on the kind of bibliometric study being conducted, co-authorship, co-citation, or co-occurrence are some of the elements that establish the network's underlying structure, which is the basis for this classification. In a network visualisation, overlapping colours are used to show occasions where elements (such as authors, articles, or keywords) are concurrently part of several clusters or communities. Colours that overlap indicate relationships or shared traits between objects in the network that cut over various theme categories.

This subsection provides a citation analysis from four different angles in order to illustrate the publications' influence. The emphasis is on showcasing the publications that are most frequently cited and offering examples of citations from different authors, organisations, and nations/regions.

4.1. Top-Ranking Significant Nations or Areas

A total of 51 countries or regions were identified, and their collaboration capacity was assessed using VOSviewer, with each location producing at least two research outputs. Figure 8 visually represents these data, where the width of the links and the connections among nodes indicate the degree and level of collaboration, respectively. The total link strength (TLS) of a node is determined by the sum of all link strengths associated with it.

In Table 3 The top ten countries are displayed according to the number of connections, which indicates the total number of countries or areas that are mentioned. The People's Republic of China, the most productive nation, is also the one that obtains the most citations. In particular, an average of 7.8968 citations are found for each Chinese publication, suggesting that these works are frequently included in academic publications. India has an impressive average of 4.4310 citations per piece as well. In terms of average citations, the USA (12.22), England (4.75), Italy (15.2), and Canada (9.0) are noteworthy countries.

Table 3. The top referenced nations or areas.

Rank	Label	TLS	P	C	AC
1	Peoples R China	48	54	660	7.90
2	USA	35	126	195	12.22
3	India	33	36	168	4.43
4	South Korea	32	116	514	5.90
5	Spain	24	20	95	6.95
6	England	19	19	70	4.75
7	Pakistan	19	20	118	3.69
8	Italy	15	22	153	15.20
9	Australia	13	13	153	8.50
10	Canada	12	22	65	9.00

India comes in second place in terms of citations with just 116 published articles, after the People's Republic of China with 126 published papers. The USA has an average of 12.22 citations per manuscript published, which is more than the People's Republic of China, which averages 7.8968. "LexRank: Graph-based lexical centrality as salience in text summarisation" [34] by Erkan, G. and Radev, DR. was published in 2004 by the USA and has the highest number of citations (1436 total).

4.2. The Most Significant Organisations

Figure 9 displays the organisation's cooperative capacity as represented by VOSviewer. The links between them make up the majority of the figure. The size of the node and the thickness of the link indicate the cooperation numbers and link strength, respectively.

There are over 56 institutions that have contributed to the research, each having at least three documents and at least four citations, as shown in Figure 9. The institutional-level citation network is examined in this section. Furthermore, the relationship between

citations and organisations for ten institutions, ten of which have earned more than forty citations, is depicted in Figure 9.

The eight clusters that comprise these 56 universities are identified by distinct colours. The line joining the two nodes indicates the associated institutions whose publications are mentioned together. The thickness of the link indicates the strength of the co-cited relationship. The larger the node, the more citations the institution's papers have received. To improve understanding, Table 4 lists the top 15 colleges with over 823 citations, along with pertinent data.

Most of the institutions displayed in Table 4 originate from India and the People's Republic of China.

Table 4. The universities with the highest citation counts.

Rank	Institute/Affiliation	Country	TLS	P	C	AC
1	Chinese Acad Sci	Peoples R China	41	18	132	7.33
2	Univ Chinese Acad Sci	Peoples R China	25	10	74	7.40
3	Univ extremadura	Spain	24	6	68	11.33
4	Sidi Mohamed Ben Abdellah Univ	Morocco	13	5	63	12.60
5	South China Univ Technol	Peoples R China	11	3	55	18.33
6	Indian Inst Informat Technol Design and Mfg	India	19	4	53	13.25
7	Univ Malaya	Malaysia	8	3	53	17.66
8	King Saud Univ	Saudi Arabia	15	6	46	7.66
9	Jawaharlal Nehru Univ	India	14	4	42	10.25
10	Princess Sumaya Univ Technol	Jordan	15	4	41	10.25
11	Beijing Univ Posts and Telecommun	Peoples R China	6	3	41	13.66
12	Islamic Azad Univ	Iran	21	6	40	6.66
13	Ibn Tofail Univ	Morocco	6	3	39	13.00
14	Univ Toulon and Var	France	6	3	39	13.00
15	Guru Gobind Singh Indraprastha Univ	India	14	3	37	12.33

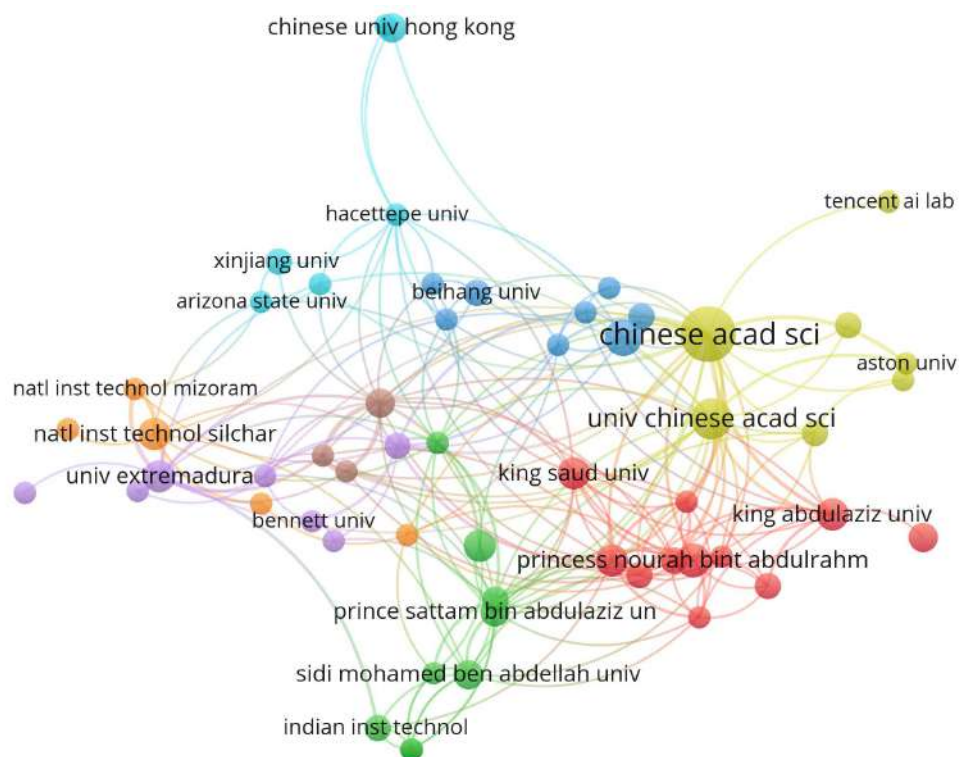


Figure 9. The citation network of institutions.

4.3. Strength of Collaboration Between the Researchers

The following describes the relationship between the co-cited and referred authors: Contributions were made at first by 106,098 authors and researchers. The results were then further refined by taking into account the author’s minimal amount of citations. Only nine authors out of the original 12,028 had at least 100 citations in this field, according to an analysis. Each of these writers received over 100 citations. The network diagram for these authors is shown in Figure 10. The tool’s default setting calls for at least 20 citations to be included; however, in order to improve the graph’s readability, we changed the threshold to 80. As a consequence, 156 writers satisfied the criteria and are shown in the figure.

Determining the degree and extent of collaboration between scholars from around the world who contribute to the research was made possible by the VOSviewer. This utility generates easily understood numerical output. Additionally, more than thirty researchers have a TLS of more than 1000, and five out of the 156 researchers with at least 20 citations each have a vast collaboration (cooperation) network with a TLS of more than 3000. It is aptly illustrated in Figure 10.

The scientific community has created 12,752 TLS and 2223 citations, as shown in Figure 10. The top researchers with the greatest TLS values are listed in Table 5.

Table 5. Top researchers with TLS scores that are the highest.

Rank	Authors	TLS	C
1	Nallapati, R	1565	252
2	Lin, CY	1513	284
3	Liu, Y	1144	187
4	See, A	953	154
5	Mihalcea, R	905	168
6	Devlin, J	881	170
7	Vaswani, A	844	147
8	Erkan, G	789	131
9	Mikolov, T	611	109
10	Luhn, HP	603	98

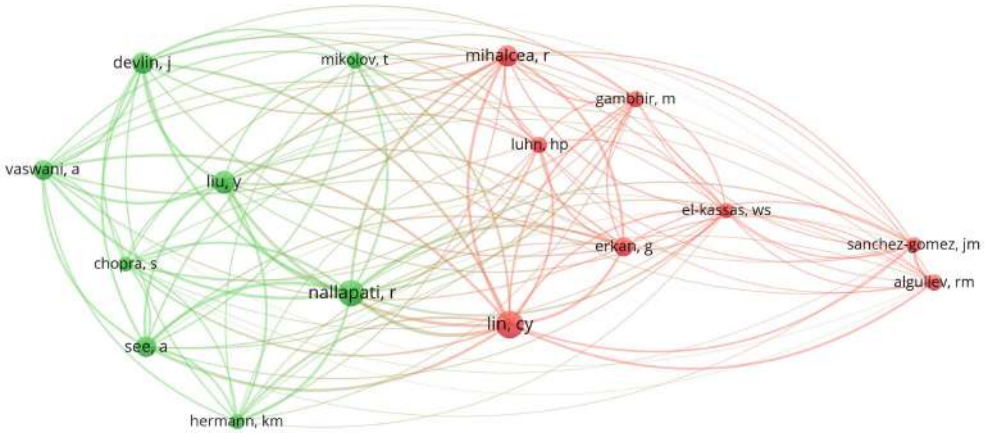


Figure 10. The citation network of authors.

4.4. Co-Citations Network of Articles

As seen in Figure 11, 14 papers obtain more than 50 citations. The figure’s nodes are sized according to how many times a publication is cited, and these citations are all combined. Three clusters have been identified, each represented by a distinct colour. The most frequently cited publications are highlighted in Table 6 along with pertinent data.

Key Influential Articles:

- Vaswani et al., 2017 (“Attention is All You Need” [35]): Article is highly central and densely connected, indicating its significant influence and widespread recognition in the academic community.
- Devlin et al., 2019 (“BERT:Pre-training of Deep Bidirectional Transformers for Language Understanding” [8]): Another highly influential work, shown by its strong connections with many other articles.
- Bahdanau et al., 2016 (“Neural Machine Translation by Jointly Learning to Align and Translate” [36]): With several co-citations, this paper also seems to have a lot of influence, which speaks to its significance in the machine translation community.

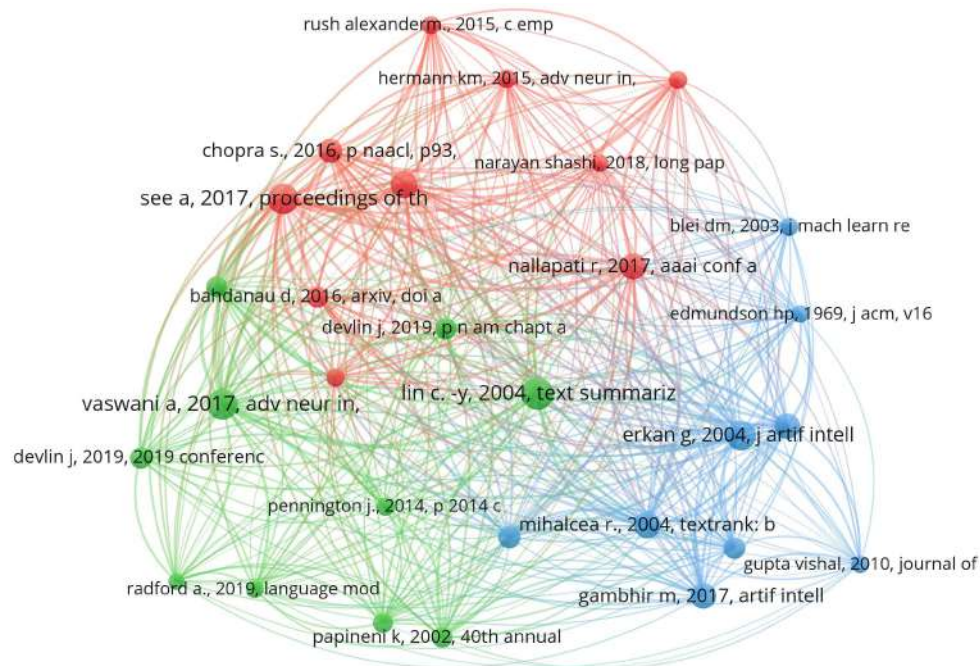


Figure 11. The article co-citation network.

4.5. Most Cited Publications

To further evaluate the features and content of the publications, a table is created with multiple variables to look at the most cited articles, the year of publication, and many more.

By ranking publications according to their influence, which is often based on the amount of citations, bibliometric analysis makes it easier to identify foundational works that have had a significant impact on the field. The inclusion of publication kinds, study focuses, and author names provides a multifaceted understanding of the types and diversity of significant contributions. Additionally, the connections among the entities aid in network analysis by demonstrating how different components of the academic discourse are interrelated. This table, which is a part of bibliometric analysis, gives researchers insights into the intellectual structure, trends, and key participants in a particular research area, enabling them to make well-informed decisions and carry out additional research on crucial contributions.

The publications in Table 6, notably “Recent automatic text summarisation techniques: a survey” [14], “A Survey on Deep Learning for Named Entity Recognition” [37], and “Automatic text summarisation: A comprehensive survey” [15], are extensively recognised studies that offer thorough insights into a range of text summarisation topics. Their appeal stems from their thorough summaries, which help to provide a comprehensive picture of the environment in which summaries are generated.

Ten publications from a variety of topics, including information systems, artificial intelligence, electrical engineering, electronics and computer science, and more in our 983 document dataset of H-index 60, are included in Table 6, indicating a considerable level of study interest. Furthermore, it is clear that there are links to other branches of computer science, including Natural Language Processing, Knowledge Engineering and Representation, and Cardiac and Cardiovascular Systems.

Table 6. Top Most Important Publications.

Rank	Article	Author	Type	Year	Research Direction	Citations	Links
1	LexRank: Graph-based lexical centrality as salience in text summarisation [34]	Erkan, G and Radev, DR	Article	2004	Computer Science, Artificial Intelligence	1436	37
2	The Structure and Dynamics of Cocitation Clusters: A Multiple-Perspective Cocitation Analysis [38]	Chen, CM; Ibekwe-SanJuan, F and Hou, JH	Article	2010	Computer Science, Information Systems, Information Science	1025	74
3	A Survey on Deep Learning for Named Entity Recognition [37]	Li, J; Sun, AX; Han, JL; Li, CL	Article	2022	Computer Science, Artificial Intelligence, Information Systems, Electrical Engineering	454	210
4	Recent automatic text summarisation techniques: A survey [14]	Gambhir, M and Gupta, V	Article	2017	Computer Science, Artificial Intelligence	355	182
5	Risk stratification for sudden cardiac death: Current status and challenges for the future [39]	Wellens, HJJ; Schwartz, PJ; Lindemans, FW; Buxton, AE; Goldberger, JJ; Hohnloser, SH; et al.	Article	2014	Cardiac & Cardiovascular Systems	284	121
6	Fitting neurological protein aggregation kinetic data via a 2-step, Minimal/“Ockham’s Razor” model: The Finke-Watzky mechanism of nucleation followed by autocatalytic surface growth [40]	Morris, AM; Watzky, MA; Agar, JN; Finke, RG	Article	2008	Biochemistry & Molecular Biology	245	67
7	Automatic text summarisation: A comprehensive survey [15]	El-Kassas, WS; Salama, CR; Rafea, AA; Mohamed, HK	Article	2021	Computer Science, Artificial Intelligence, Electrical Engineering, Operations Research	233	227
8	An overview of the BIOASQ large-scale biomedical semantic indexing and question answering competition [41]	Tsatsaronis, G; et al.	Article	2015	Biochemical Research, Biotechnology, Mathematical & Computational Biology	233	37
9	Opinion mining from online hotel reviews—A text summarisation approach [42]	Hu, YH; Chen, YL and Chou, HL	Article	2017	Computer Science, Information Systems, Information Science	184	61
10	Assessing sentence scoring techniques for extractive text summarisation [43]	Ferreira, R; Cabral, LDS; Lins, RD; Silva, GPE; et al.	Article	2013	Computer Science, Artificial Intelligence, Electrical Engineering, Operations Research	173	35

4.6. Analysis of Cooperation

Collaboration networks between countries/regions, institutions, and authors are built using the VOSviewer software, which facilitates the evaluation of collaborative links and promotes a more profound comprehension of research collaborations in the field. Relevant data were extracted from WoS and imported into VOSviewer to analyse regional and national cooperation networks; 39 of the 65 countries or regions that reached the requirement and created collaborative networks did so by requiring a minimum of two papers, as shown in Figure 12.

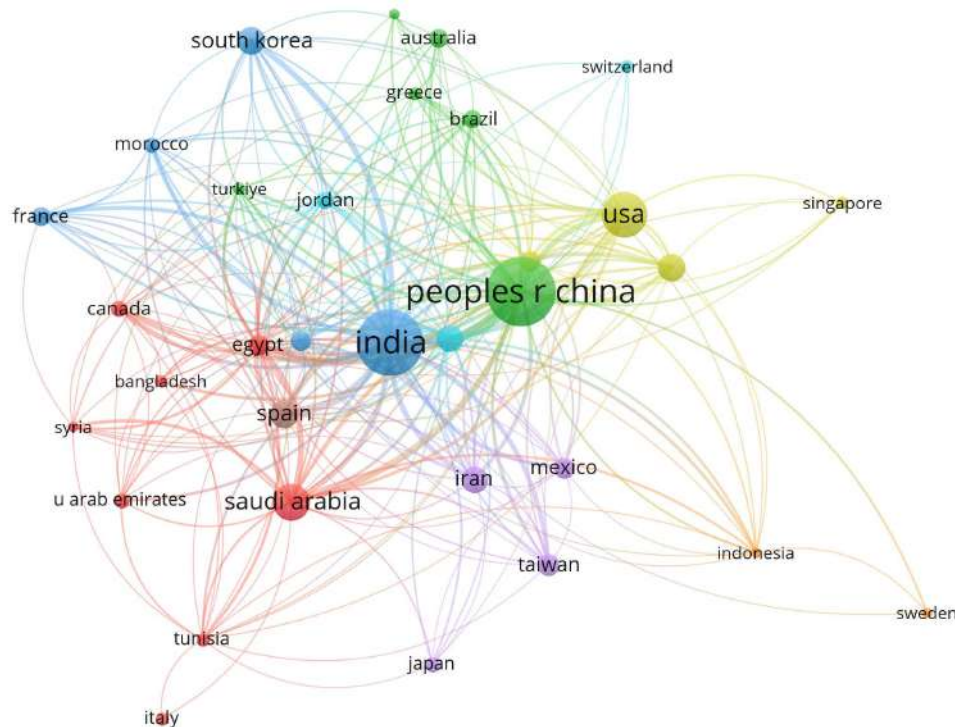


Figure 12. The citation network of places.

Ten clusters representing different cooperative ties are created by further classifying these 39 countries and areas, each identified distinctly. Collaborations are represented by the connections between nodes, and the strength or frequency of cooperation is shown by the breadth of the link. A node's size is determined by its TLS, which is the total of all link strengths.

The top three countries with the highest TLS in Figure 12 are Egypt, India, and the People's Republic of China, indicating their considerable cooperation efforts.

5. Publishers over the Years

In this section, we will be discussing some trends regarding publishers over the years. A chord diagram was made using Raw Graphs, an online graph tool, to represent a measure of publishers publishing in a specific year.

The given graph in Figure 13 shows the top 18 publishers from 2010 to 2024 who published at least one paper in their respective target years. Here are some trends and observations:

- **Prominent Publishers:** Springer and Elsevier are the most prominent publishers, indicated by their strong connections over multiple years. They have consistently published works across a wide range of years, demonstrating their significant presence in academic publishing.
- **Recent Publications:** IEEE Computer Society, Springer, and Elsevier show substantial activity in recent years (2021–2024). This indicates a trend where these publishers are heavily involved in contemporary academic and scientific literature.
- **Emerging Publishers:** MDPI and PeerJ Inc have been connected for multiple years, suggesting they are emerging as notable publishers in recent times. Their increasing connections indicate growing influence and publication volume.
- **Diverse Publishers:** Publishers like Springer Nature and Cambridge University Press have connections across a wide range of years, indicating their long-standing involvement in academic publishing.

- **Consistent Contributors:** Publishers such as Elsevier Science Inc, Springer Heidelberg, and IEEE Computer Society show consistent publication activity over the years, suggesting a stable and ongoing contribution to the academic and scientific community.
- **Specialised Publishers:** Some publishers have a focused presence in specific years, such as Academic Press Ltd in earlier years and Nature Portfolio in more recent years. This may indicate periods of heightened activity or focus on particular research areas during those times.
- **Publication Peaks:** Certain years, particularly 2023 and 2024, show high publication activity across multiple publishers, suggesting recent years have seen a surge in academic output.

These trends illustrate the dynamic landscape of academic publishing, with major players maintaining their influence while new and specialised publishers are emerging and contributing significantly to the field.

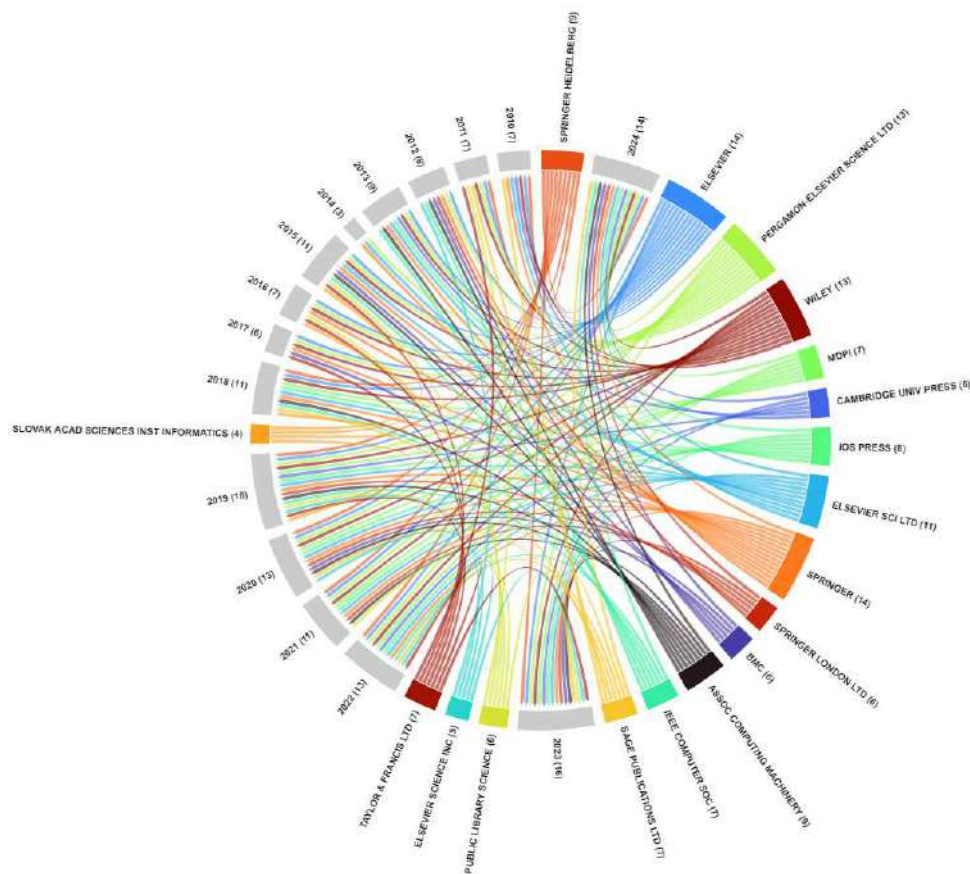


Figure 13. Chord diagram of publisher and publishing year.

6. Survey Findings and Future Directions

These sections explore the current status and future directions of literature search in various ways. A summary of text summarisation followed by a literature survey discussing strategies, key features, and limitations is provided in Table 7. Additionally, the working of the model is explained by outlining the processes of several papers as shown in Table 8, along with a flow diagram illustrating the overall process.

The fundamental differences between abstractive and extractive summarisation methods, as discussed earlier in the introduction, are now elaborated with a focus on the specific

techniques used in each category. To provide a distinction, the methods have been classified into abstractive and extractive approaches, which are illustrated in Figure 14. This figure highlights the distinct processes and applications of these methods in generating concise textual representations. For extractive summarisation, the focus is on selecting and assembling key sentences or phrases directly from the source text, employing techniques such as statistical methods, traditional machine learning, and neural models. Conversely, abstractive summarisation generates new sentences that paraphrase or synthesize content from the source text, leveraging advanced methods including sequence-to-sequence models, transformers, and other deep learning frameworks.

These methods are also referenced in the summary of text summarisation techniques, where their overall role and classification in the field are mentioned.

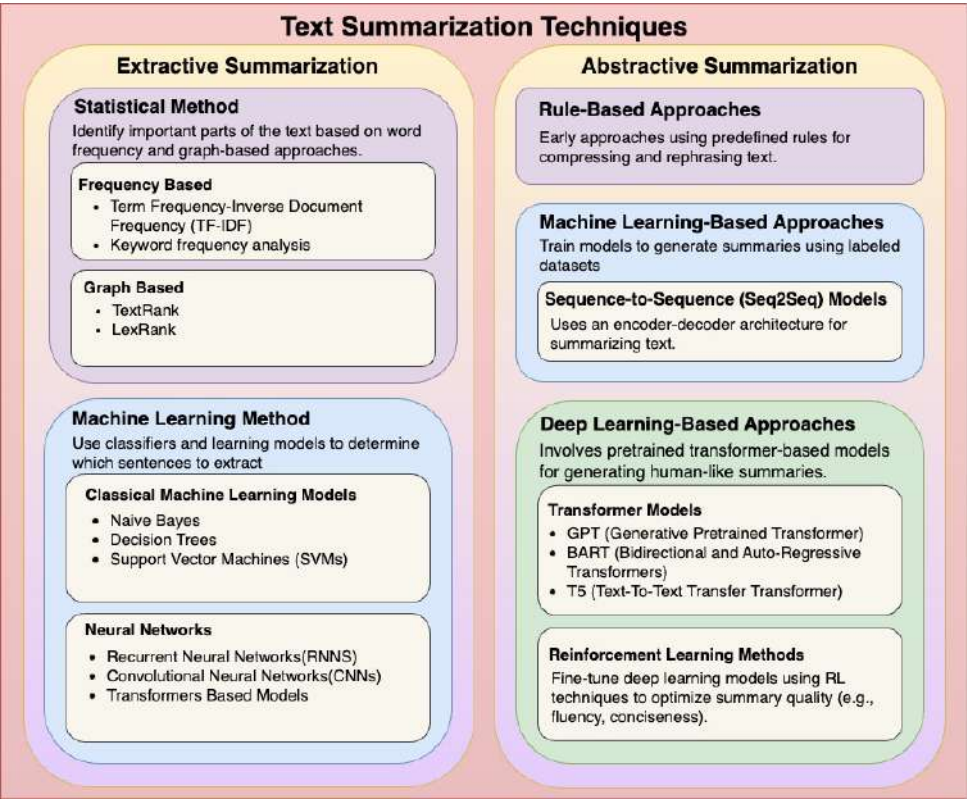


Figure 14. An overview of Text Summarisation Techniques.

Table 7 provides a concise overview of important developments in text summarisation techniques, highlighting notable contributions from different authors throughout the years. The document details the authors along with their publication years, the main objectives of their research, the methodologies employed, and the significant characteristics of each approach. The table also delineates the limitations recognised in each major paper, providing a comprehensive overview of the developing methodologies and challenges within the domain of text summarisation. Additional information regarding significant advancements can be found in Table 8. The structure and functioning of the model are described by detailing the methodologies presented in various papers, followed by a flow diagram in Appendix A that represents the comprehensive process. In diagram Appendix A the box represents one computation unit and the arrow represents the flow of the process.

Table 7. Summary of key advancements in text summarisation techniques.

Author & Year	Aim	Strategy	Key Feature	Limitation
Sascha Rothe, Shashi Narayan, Aliaksei Severyn [44], 2020	Leveraging Pre-trained Checkpoints for Sequence Generation Tasks	Pre-trained models, fine-tuning, transfer learning	Adapt pre-trained models to specific tasks. Improves task performance	Pre-trained dependency, fine-tuning needs, domain mismatch, resource-intensive
Iz Beltagy, Matthew E. Peters, Arman Cohan [45], 2020	Longformer: The Long-Document Transformer	Efficient attention, long-document handling, transformer adaptation	Local attention, handles long documents, reduces memory requirements	Complex implementation, resource-intensive
Jingqing Zhang, Yao Zhao, Mohammad Saleh, Peter J. Liu [13], 2020	PEGASUS: Pre-training with Extracted Gap-sentences	Gap-sentence generation, pre-training, fine-tuning	Masks and predicts whole sentences, concise summaries	Data-heavy, computationally expensive
Mike Lewis et al. [46], 2020	BART: Denoising Sequence-to-Sequence Pre-training	Denoising pre-training, encoder-decoder architecture	Corrupts and reconstructs text for robustness, versatile applications	Computational cost, data requirements
Yang Liu, Mirella Lapata [47], 2019	Text summarisation with Pretrained Encoders	Pretrained encoders, sequence-to-sequence model	Utilizes BERT for summarisation tasks, sequence-to-sequence approach	Computational cost, fine-tuning challenges
Abigail See, Peter J. Liu, Christopher D. Manning [48], 2017	Get to the Point: summarisation with Pointer-Generator Networks	Pointer-generator network, coverage mechanism	Combines generation and copying from original text	High computational requirements, repetition issues
Alexander M. Rush, Sumit Chopra, Jason Weston [9], 2015	A Neural Attention Model for Abstractive summarisation	Neural network, attention mechanism, encoder-decoder framework	Focuses attention on key input parts for abstractive summarisation	Data dependence, complexity, resource-intensive
Barzilay, R., Elhadad, M. [49], 2003	Sentence Alignment for Monolingual Comparable Corpora	Sentence alignment, similarity measures, monolingual corpora	Aligns sentences based on similarity measures in related content	Scalability, alignment accuracy, complexity
Mani, I., Bloedorn, E., Gates, B. [50], 2002	Machine Learning of Generic and User-Focused summarisation	Feature extraction, machine learning models, summarisation	Identifies relevant features for summarisation, machine learning-based	Data dependence, feature selection, scalability
Lin, C.-Y. [51], 2004	ROUGE: A Package for Automatic Evaluation of Summaries	Evaluation metrics, n-gram overlap	Precision, recall, F1 score for comparing summaries	Surface-level analysis, limited semantic understanding
Nenkova, A., Vanderwende, L. [52], 2005	The Impact of Frequency on summarisation	Frequency analysis, sentence scoring, extraction	Ranks sentences by term frequency, compares frequency-based methods	Context ignorance, domain sensitivity
Erkan, G., Radev, D.R. [34], 2004	LexRank: Graph-based Lexical Centrality	Graph-based ranking, LexRank algorithm, sentence selection	Uses a graph for sentence ranking based on lexical similarity	Computational complexity, domain sensitivity
Mihalcea, R., Tarau, P. [6], 2004	TextRank: Bringing Order into Texts	Graph-based ranking, TextRank algorithms	Uses PageRank to rank sentences in a graph based on similarity	Computationally intensive, context ignorance
Gong, Y., Liu, X. [3], 2001	Generic summarisation Using Latent Semantic Analysis	Relevance measure, latent semantic analysis	Uses LSA for dimensionality reduction, selects key sentences	Context ignorance, computational complexity
Eduard Hovy, Chin-Yew Lin [53], 2000	Automated Acquisition of Topic Signatures	Topic signatures, sentence scoring, summary extraction	Acquires key terms for topics, ranks sentences based on them	Context ignorance, relevance overlap, complexity
Julian Kupiec et al. [54], 1995	A Trainable Document Summarizer	Trainable summarisation, feature extraction	Features include term frequency, sentence position, and cue phrases	Domain dependence, requires a large training corpus
Harold P. Edmundson [11], 1969	New Methods in Automatic Extracting	Statistical analysis, keyword weighting, human editing	Sentence position importance, keyword relevance	Limited scope, reliance on sentence position
H.P. Luhn [10], 1958	The Automatic Creation of Literature Abstracts	Word frequency, sentence scoring, selection criteria	Frequency analysis, content scoring, human editing	Simplistic scoring, manual intervention

Table 8. Summary of models and their methodologies for text summarisation.

Figure and Model	Input	Key Methodology	Key Features	Summary Generation Process
Figure A1a Luhn [10]	Large Text	Frequency and position-based methods	Statistical methods, Term frequency	Selects significant words and generates summary
Figure A1b Edmundson [11]	Large Text	Frequency, pragmatic words, and structural indicators	Hybrid approach (linguistic & structural features)	Selects significant sentences
Figure A1c Kupiec et al. [54]	Large Text	Sentence scoring, Bayes' theorem	Supervised learning, Term frequency, Cue words	Chooses the most probable sentences based on Bayes' probability
Figure A1d Lin and Hovy [53]	Large Text	Topic signature and sentence ranking	Semantic relevance, Corpus-based methods	Ranks sentences and selects the top-ranked ones
Figure A1e Gong and Liu [3]	Matrix form	Singular Value Decomposition (SVD)	Dimensionality reduction, Semantic space projection	Identifies key sentences using SVD
Figure A1f Mihalcea and Tarau [6]	Large Text	TextRank Graph	Unsupervised learning, Sentence similarity	Selects sentences from the graph
Figure A1g Erkan and Radev [34]	Large Text	LexRank Graph	Lexical analysis, Sentence connectivity	Selects sentences from the graph
Figure A1h Nenkova et al. [52]	Large Text	Frequency-based method	Sentence probability revision	Selects highest-probability sentences iteratively
Figure A1i Mani and Bloedorn [50]	Large Text	SCDF learning algorithms	Adaptable learning rules	Customizes rules to generate summaries
Figure A1j Barzilay and Elhadad [49]	Large Text	Structural alignment of sentences	Macro and Micro alignment	Aligns and generates summaries
Figure A1k Rush et al. [9]	Large Text	Encoder-Decoder model	Attention mechanism, Beam search decoder	Generates summaries with state-of-the-art results
Figure A1l See et al. [48]	Large Text	LSTM Bi-directional Encoder	Attention mechanism or Unidirectional Decoder	Generates summaries effectively
Figure A1m Liu and Lapata [47]	Large Text	BERT with internal segmentation embeddings	Pre-trained encoder	Generates improved summaries
Figure A1n Lewis et al. [46]	Large Text	Random and pre-trained encoder-decoder	Transfer learning for fine-tuning	Improves summarisation, machine translation, and Q&A tasks
Figure A1o Zhang et al. [13]	Large Text	Transformer encoder-decoder	Task-specific pre-training objectives	Generates high-quality summaries
Figure A1p Beltagy et al. [45]	Large Text	Transformer Encoder-Decoder with Longformer	Local and global attention mechanisms	Reduces computational overhead for long-context tasks like document summarisation, improving efficiency and scalability.
Figure A1q Rothe et al. [44]	Large Text	Pre-trained BERT/RoBERTa/GPT-2	Transfer learning	Improves summarisation outputs

6.1. Text Summarisation Frameworks

This section presents a block diagram approach to illustrate the frameworks proposed in various research papers on text summarisation. The diagrams are discussed in detail and visually represented using block diagrams for a clearer understanding of the summarisation processes. In the figure, each model leverages distinct methodologies to process large pieces of text and generate concise summaries. The diagrams provide a visual representation of inputs, processes, and outputs for each model, facilitating a clearer understanding of their unique approaches. By delineating these methodologies, we aim to highlight the diversity of strategies available in text summarisation, providing insight into how each approach can be applied to enhance our data processing and summarisation objectives.

6.2. Future Directions

The findings of this bibliometric analysis reveal several promising avenues for future research in the field of text summarisation. Based on the identified trends, challenges, and gaps, the following directions are proposed

- **Improved Abstractive Summarisation**—Develop more coherent and semantically accurate abstractive models [55] that generate summaries closer to human-written content. Focus on reducing issues such as factual inconsistencies and hallucinations in summaries by integrating external knowledge bases.
- **Domain-Specific Summarisation**—Create customised summarisation systems for specialised fields such as healthcare, law, and finance. Emphasize the development of domain-specific datasets and pre-trained models tailored to industry jargon and technical content [56].
- **Real-Time Summarisation**—Design efficient models capable of summarising streaming data from sources [57] like social media, live news feeds, and real-time analytics. Incorporate temporal dynamics and event detection to enhance the relevance and timeliness of the generated summaries.
- **Multimodal Summarisation**—Focus on building models that integrate audio, video, images, multiple languages [17], and other types of data. Develop systems that extract and generate important information from different forms of data.
- **Dialogue Summarisation**—Improve summarisation techniques for multi-turn conversations [58], such as those in meetings, interviews, or customer service interactions. Address challenges like preserving speaker-specific information, maintaining coherence, and summarising complex dialogue structures.
- **Bias and Ethics**—Mitigate biases [59] inherent in training datasets that can lead to skewed or unfair summaries. Develop ethical guidelines and frameworks for creating transparent, fair, and accountable summarisation systems. Introduce tools to detect and rectify biased summaries in real time.
- **Multilingual Summarisation**—Extend summarisation capabilities to support multiple languages [17], especially those with limited resources. Leverage transfer learning and multilingual transformer models to build systems adaptable to diverse linguistic contexts.
- **Interactive Summarisation**—Develop interactive systems that allow users to customize summaries by specifying preferences, such as length, focus, or tone [60]. Incorporate feedback mechanisms to improve summary quality based on user interactions.
- **Low-Resource Summarisation**—Address the lack of large datasets for low-resource languages or domains by exploring data augmentation techniques and unsupervised learning [61]. Utilize transfer learning and cross-lingual approaches to adapt high-resource models for low-resource settings.
- **Explainability**—Enhance model interpretability [62] by providing insights into why specific sentences or phrases were selected for the summary. Develop visualisation tools that highlight attention weights and decision pathways in the summarisation process.
- **Long Document Summarisation**—Design models optimised for processing and summarising lengthy documents, such as books, legal texts, and research articles. Focus on improving memory efficiency [63] and coherence over extended content using techniques like hierarchical attention.
- **Better Evaluation Metrics**—Move beyond traditional metrics like ROUGE and BLEU to develop comprehensive, human-centric evaluation frameworks. Incorporate measures of semantic accuracy, readability, coherence, and task-specific performance for a holistic assessment [64].

Text summarisation research has significant implications for information processing, but ethical concerns need to be addressed as technology evolves. Key issues include ensuring accuracy to avoid misrepresentation, mitigating bias to ensure fairness, protect-

ing privacy through anonymisation, and obtaining consent for data usage. Transparency in methods and accountability for model outputs are essential to build trust. Furthermore, the societal impact of summarisation technologies should be considered, particularly in preventing misuse and ensuring accessibility for all users. Future research should focus on refining models to address these ethical challenges while advancing summarisation techniques.

Limitations

While this study provides a comprehensive analysis of text summarisation research, it has some limitations. The analysis relies solely on publications indexed in the WoS database, which may exclude relevant studies from other sources. Only English-language publications were included, potentially missing valuable contributions in other languages. Insights from this study may not generalize to all fields or domains beyond text summarisation. Addressing these limitations in future work could enhance the robustness of the findings.

7. Conclusions

This study conducts a bibliometric analysis of research and survey papers published between 2004 and September 2024 to recognize developments in text summarisation research. The field has gained significant popularity, boasting an H-index of 60, with 2023 emerging as the year with the highest number of publications. The analysis reveals a steady growth in the number of publications over the years, driven by significant contributions from countries such as China, India, and the USA. Notably, China leads with 201 publications and a Total Link Strength (TLS) of 48. Future growth in research output is anticipated. Prominent journals in this domain include Expert Systems with Applications and IEEE Access. Institutions such as the Chinese Academy of Sciences, the Indian Institute of Technology, and King Saud University are highly active contributors. Among the authors, Lloret E. has published the most articles, with other researchers playing key roles in driving innovation. The study highlights the transition from traditional extractive methods, like TextRank, to advanced transformer-based abstractive models, such as BERT and GPT. Future directions for the field include topics like multimodal and multilingual summarisation. Additionally, the review offers valuable insights into semantic analysis techniques, complemented by two comprehensive tables and block diagrams that succinctly illustrate significant advancements and research gaps in the field. This work significantly contributes to understanding and advancing text summarisation methodologies.

Author Contributions: Methodology, A.K.; Resources, C.-C.L.; Writing—original draft, R.C.; Writing—review & editing, R.K.; Funding acquisition, C.-C.L. All authors have read and agreed to the published version of the manuscript.

Funding: This National Science and Technology Council under Grant (NSC 111-2410-H-167-005-MY2 and NSC 113-2410-H-167-012-MY3).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Email requests will be answered regarding the data utilised in the studies in this research project.

Conflicts of Interest: The authors have no conflicting agendas.

Appendix A. Advancements in Text Summarisation Models

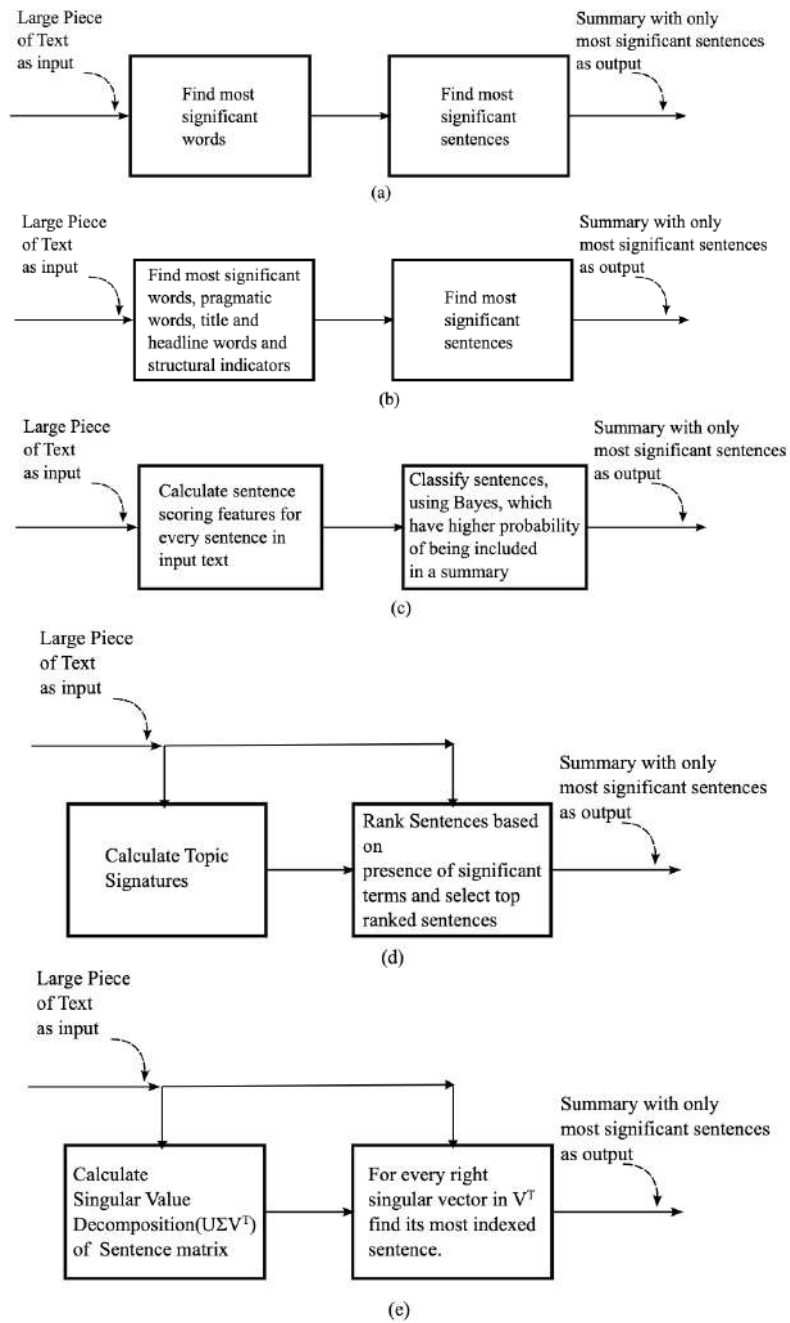


Figure A1. Cont.

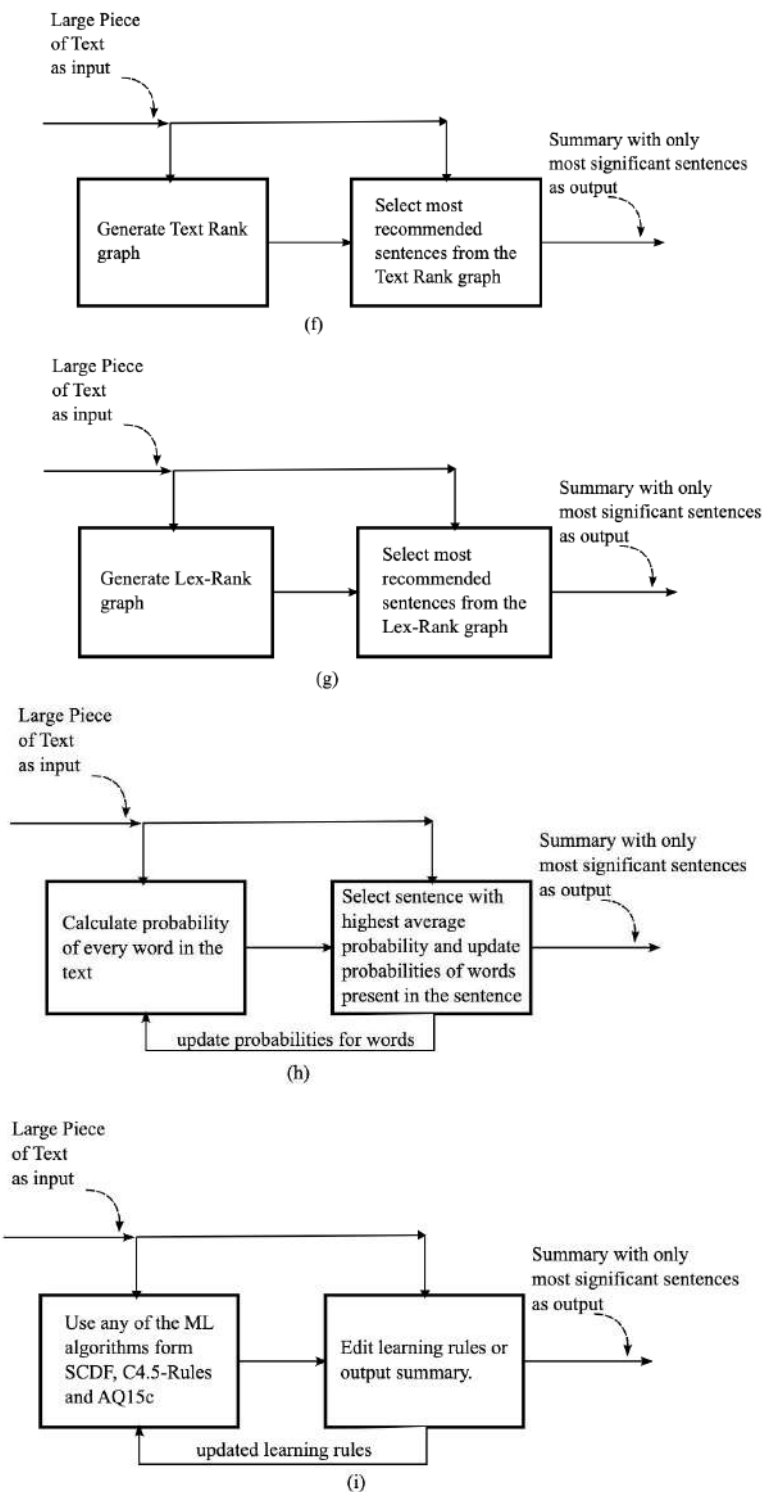


Figure A1. Cont.

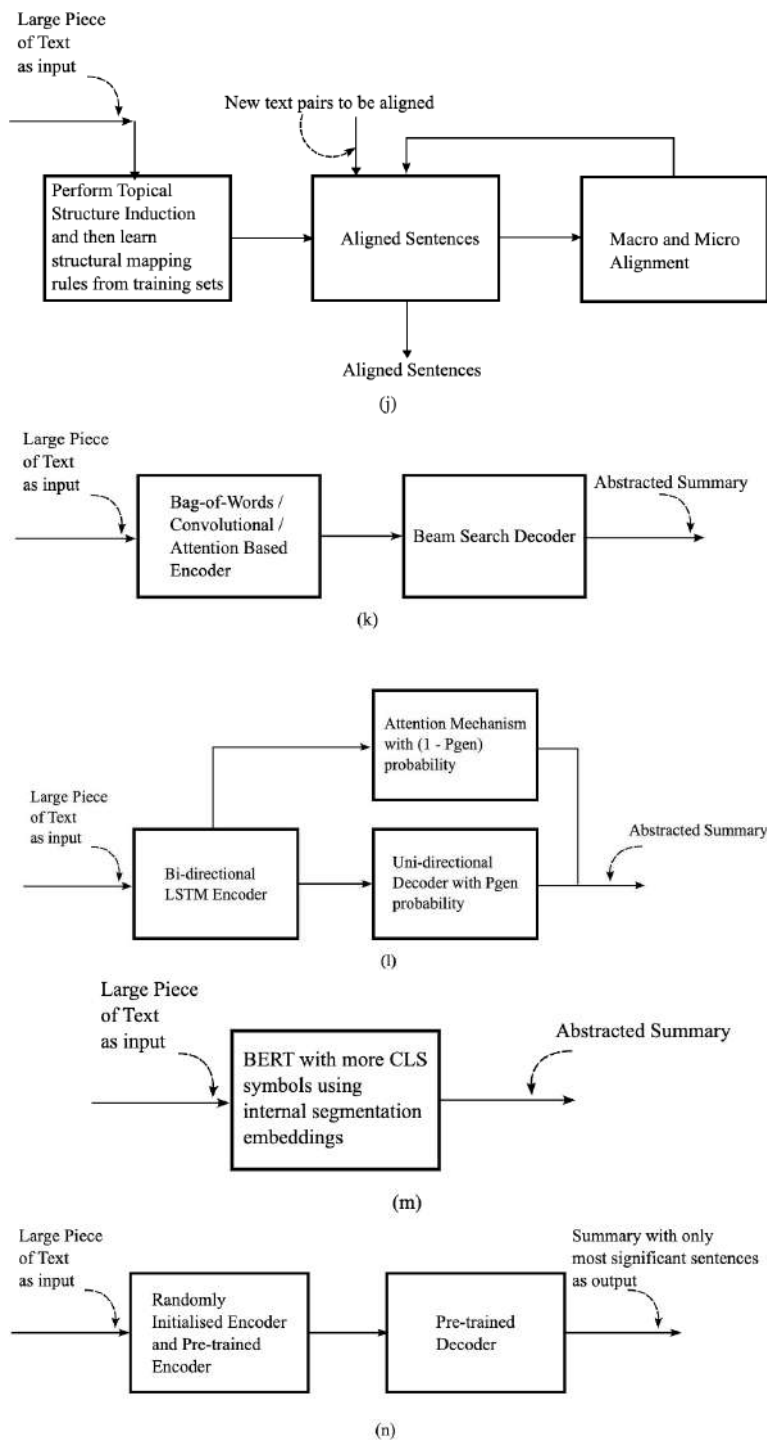


Figure A1. Cont.

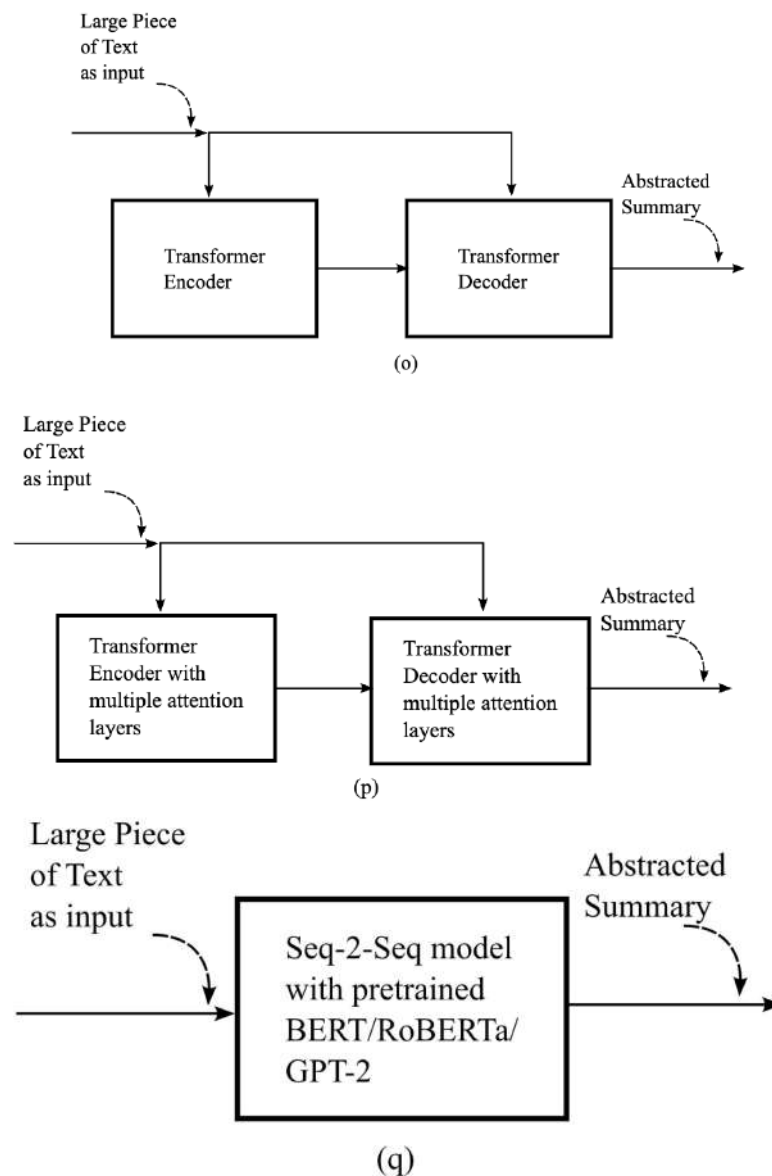


Figure A1. Block diagrams illustrating the processes of significant research papers on text summarisation.

References

1. Malik, N.; Bilal, M. Natural language processing for analyzing online customer reviews: A survey, taxonomy, and open research challenges. *PeerJ Comput. Sci.* **2024**, *10*, e2203. [\[CrossRef\]](#) [\[PubMed\]](#)
2. Min, B.; Ross, H.; Sulem, E.; Veyseh, A.P.B.; Nguyen, T.H.; Sainz, O.; Agirre, E.; Heintz, I.; Roth, D. Recent advances in natural language processing via large pre-trained language models: A survey. *ACM Comput. Surv.* **2023**, *56*, 1–40. [\[CrossRef\]](#)
3. Gong, Y.; Liu, X. Generic text summarization using relevance measure and latent semantic analysis. In Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '01), New York, NY, USA, 9–13 September 2001; pp. 19–25. [\[CrossRef\]](#)
4. Luhn, H.P. A Statistical Approach to Mechanized Encoding and Searching of Literary Information. *IBM J. Res. Dev.* **1957**, *1*, 309–317. [\[CrossRef\]](#)
5. Gupta, V.; Lehal, G.S. A survey of text summarization extractive techniques. *J. Emerg. Technol. Web Intell.* **2010**, *2*, 258–268. [\[CrossRef\]](#)
6. Mihalcea, R.; Tarau, P. TextRank: Bringing order into text. In Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, Barcelona, Spain, 25–26 July 2004; pp. 404–411.

7. Goudjil, M.; Koudil, M.; Bedda, M.; Ghoggali, N. A Novel Active Learning Method Using SVM for Text Classification. *Int. J. Autom. Comput.* **2016**, *15*, 290–298. [\[CrossRef\]](#)
8. Devlin, J. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
9. Rush, A.M.; Chopra, S.; Weston, J. A Neural Attention Model for Abstractive Sentence Summarization. *arXiv* **2015**, arXiv:1509.00685. [\[CrossRef\]](#)
10. Luhn, H.P. The Automatic Creation of Literature Abstracts. *IBM J. Res. Dev.* **1958**, *2*, 159–165. [\[CrossRef\]](#)
11. Edmundson, H.P. New Methods in Automatic Extracting. *J. ACM* **1969**, *16*, 264–285. [\[CrossRef\]](#)
12. Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I. Language models are unsupervised multitask learners. *OpenAI Blog* **2019**, *1*, 9.
13. Zhang, J.; Zhao, Y.; Saleh, M.; Liu, P. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In Proceedings of the International Conference on Machine Learning, PMLR, Virtual, 13–18 July 2020; pp. 11328–11339.
14. Gambhir, M.; Gupta, V. Recent automatic text summarization techniques: A survey. *Artif. Intell. Rev.* **2017**, *47*, 1–66. [\[CrossRef\]](#)
15. El-Kassas, W.S.; Salama, C.R.; Rafea, A.A.; Mohamed, H.K. Automatic text summarization: A comprehensive survey. *Expert Syst. Appl.* **2021**, *165*, 113679. [\[CrossRef\]](#)
16. Chen, J.; Zhuge, H. Abstractive Text-Image Summarization Using Multi-Modal Attentional Hierarchical RNN. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, 31 October–4 November 2018; Riloff, E., Chiang, D., Hockenmaier, J., Tsujii, J., Eds.; pp. 4046–4056. [\[CrossRef\]](#)
17. Taunk, D.; Varma, V. Summarizing Indian Languages using Multilingual Transformers based Models. *arXiv* **2023**, arXiv:2303.16657. [\[CrossRef\]](#)
18. Chandler, A.; Surve, D.; Su, H. Detecting Errors through Ensembling Prompts (DEEP): An End-to-End LLM Framework for Detecting Factual Errors. *arXiv* **2024**, arXiv:2406.13009. [\[CrossRef\]](#)
19. Xu, P.; Shao, W.; Chen, M.; Tang, S.; Zhang, K.; Gao, P.; An, F.; Qiao, Y.; Luo, P. BESA: Pruning Large Language Models with Blockwise Parameter-Efficient Sparsity Allocation. *arXiv* **2024**, arXiv:2402.16880. [\[CrossRef\]](#)
20. Zhang, H.; Song, H.; Li, S.; Zhou, M.; Song, D. A survey of controllable text generation using transformer-based pre-trained language models. *ACM Comput. Surv.* **2023**, *56*, 1–37. [\[CrossRef\]](#)
21. Jangra, A.; Mukherjee, S.; Jatowt, A.; Saha, S.; Hasanuzzaman, M. A survey on multi-modal summarization. *ACM Comput. Surv.* **2023**, *55*, 1–36. [\[CrossRef\]](#)
22. Mridha, M.F.; Lima, A.A.; Nur, K.; Das, S.C.; Hasan, M.; Kabir, M.M. A survey of automatic text summarization: Progress, process and challenges. *IEEE Access* **2021**, *9*, 156043–156070. [\[CrossRef\]](#)
23. Koh, H.Y.; Ju, J.; Liu, M.; Pan, S. An empirical survey on long document summarization: Datasets, models, and metrics. *ACM Comput. Surv.* **2022**, *55*, 1–35. [\[CrossRef\]](#)
24. Khan, B.; Shah, Z.A.; Usman, M.; Khan, I.; Niazi, B. Exploring the landscape of automatic text summarization: A comprehensive survey. *IEEE Access* **2023**, *11*, 109819–109840. [\[CrossRef\]](#)
25. Chen, X.; Xie, H.; Tao, X.; Xu, L.; Wang, J.; Dai, H.N.; Wang, F.L. A topic modeling-based bibliometric exploration of automatic summarization research. In *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*; Wiley: Hoboken, NJ, USA, 2024; p. e1540.
26. Kaushal, A.; Kumar, S.; Kumar, R. A review on deepfake generation and detection: Bibliometric analysis. In *Multimedia Tools and Applications*; Springer: Berlin/Heidelberg, Germany, 2024; pp. 1–41.
27. Ellegaard, O.; Wallin, J.A. The bibliometric analysis of scholarly production: How great is the impact? *Scientometrics* **2015**, *105*, 1809–1831. [\[CrossRef\]](#) [\[PubMed\]](#)
28. Kanapala, A.; Pal, S.; Pamula, R. Text summarization from legal documents: A survey. *Artif. Intell. Rev.* **2019**, *51*, 371–402. [\[CrossRef\]](#)
29. Saini, N.; Saha, S.; Jangra, A.; Bhattacharyya, P. Extractive single document summarization using multi-objective optimization: Exploring self-organized differential evolution, grey wolf optimizer and water cycle algorithm. *Knowl.-Based Syst.* **2019**, *164*, 45–67. [\[CrossRef\]](#)
30. Verma, P.; Om, H. MCRMR: Maximum coverage and relevancy with minimal redundancy based multi-document summarization. *Expert Syst. Appl.* **2019**, *120*, 43–56. [\[CrossRef\]](#)
31. Saini, N.; Saha, S.; Bhattacharyya, P. Automatic Scientific Document Clustering Using Self-organized Multi-objective Differential Evolution. *Cogn. Comput.* **2019**, *11*, 271–293. [\[CrossRef\]](#)
32. Lloret, E.; Palomar, M. Text summarisation in progress: A literature review. *Artif. Intell. Rev.* **2012**, *37*, 1–41. [\[CrossRef\]](#)
33. Sanchez-Gomez, J.M.; Vega-Rodriguez, M.A.; Perez, C.J. Extractive multi-document text summarization using a multi-objective artificial bee colony optimization approach. *Knowl.-Based Syst.* **2018**, *159*, 1–8. [\[CrossRef\]](#)
34. Erkan, G.; Radev, D. LexRank: Graph-based lexical centrality as salience in text summarization. *J. Artif. Intell. Res.* **2004**, *22*, 457–479. [\[CrossRef\]](#)
35. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. *arXiv* **2017**, arXiv:1706.03762. [\[CrossRef\]](#)
36. Bahdanau, D.; Cho, K.; Bengio, Y. Neural Machine Translation by Jointly Learning to Align and Translate. *arXiv* **2014**, arXiv:1409.0473.

37. Li, J.; Sun, A.; Han, J.; Li, C. A Survey on Deep Learning for Named Entity Recognition. *IEEE Trans. Knowl. Data Eng.* **2022**, *34*, 50–70. [\[CrossRef\]](#)
38. Chen, C.; Ibekwe-SanJuan, F.; Hou, J. The Structure and Dynamics of Cocitation Clusters: A Multiple-Perspective Cocitation Analysis. *J. Am. Soc. Inf. Sci. Technol.* **2010**, *61*, 1386–1409. [\[CrossRef\]](#)
39. Wellens, H.J.J.; Schwartz, P.J.; Lindemans, F.W.; Buxton, A.E.; Goldberger, J.J.; Hohnloser, S.H.; Huikuri, H.V.; Kaeae, S.; La Rovere, M.T.; Malik, M.; et al. Risk stratification for sudden cardiac death: Current status and challenges for the future. *Eur. Heart J.* **2014**, *35*, 1642–1651. [\[CrossRef\]](#) [\[PubMed\]](#)
40. Morris, A.M.; Watzky, M.A.; Agar, J.N.; Finke, R.G. Fitting neurological protein aggregation kinetic data via a 2-step, Minimal/“Ockham’s Razor” model: The Finke-Watzky mechanism of nucleation followed by autocatalytic surface growth. *Biochemistry* **2008**, *47*, 2413–2427. [\[CrossRef\]](#)
41. Tsatsaronis, G.; Balikas, G.; Malakasiotis, P.; Partalas, I.; Zschunke, M.; Alvers, M.R.; Weissenborn, D.; Krithara, A.; Petridis, S.; Polychronopoulos, D.; et al. An overview of the BIOASQ large-scale biomedical semantic indexing and question answering competition. *BMC Bioinform.* **2015**, *16*, 138. [\[CrossRef\]](#)
42. Hu, Y.H.; Chen, Y.L.; Chou, H.L. Opinion mining from online hotel reviews - A text summarization approach. *Inf. Process. Manag.* **2017**, *53*, 436–449. [\[CrossRef\]](#)
43. Ferreira, R.; Cabral, L.d.S.; Lins, R.D.; Pereira e Silva, G.; Freitas, F.; Cavalcanti, G.D.C.; Lima, R.; Simske, S.J.; Favaro, L. Assessing sentence scoring techniques for extractive text summarization. *Expert Syst. Appl.* **2013**, *40*, 5755–5764. [\[CrossRef\]](#)
44. Rothe, S.; Narayan, S.; Severyn, A. Leveraging Pre-trained Checkpoints for Sequence Generation Tasks. *Trans. Assoc. Comput. Linguist.* **2020**, *8*, 264–280. [\[CrossRef\]](#)
45. Beltagy, I.; Peters, M.E.; Cohan, A. Longformer: The Long-Document Transformer. *arXiv* **2020**, arXiv:2004.05150. [\[CrossRef\]](#)
46. Lewis, M.; Liu, Y.; Goyal, N.; Ghazvininejad, M.; Mohamed, A.; Levy, O.; Stoyanov, V.; Zettlemoyer, L. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. *arXiv* **2019**, arXiv:1910.13461. [\[CrossRef\]](#)
47. Liu, Y.; Lapata, M. Text Summarization with Pretrained Encoders. *arXiv* **2019**, arXiv:1908.08345. [\[CrossRef\]](#)
48. See, A.; Liu, P.J.; Manning, C.D. Get To The Point: Summarization with Pointer-Generator Networks. *arXiv* **2017**, arXiv:1704.04368. [\[CrossRef\]](#)
49. Barzilay, R.; Elhadad, N. Sentence Alignment for Monolingual Comparable Corpora. In Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing, Sapporo, Japan, 11–12 July 2003; pp. 25–32.
50. Mani, I.; Bloedorn, E. Machine Learning of Generic and User-Focused Summarization. *arXiv* **1998**, arXiv:cs.CL/9811006.
51. Lin, C.Y. ROUGE: A Package for Automatic Evaluation of Summaries. In Proceedings of the Text Summarization Branches Out, Barcelona, Spain, 25–26 July 2004; pp. 74–81.
52. Nenkova, A.; Vanderwende, L.; Vanderwende, L. The Impact of Frequency on Summarization. 2005. Available online: <https://www.semanticscholar.org/paper/The-Impact-of-Frequency-on-Summarization-Nenkova-Vanderwende/676b1549adae511164c1b5343f10260fd42035b4> (accessed on 30 November 2024).
53. Lin, C.Y.; Hovy, E. The Automated Acquisition of Topic Signatures for Text Summarization. In Proceedings of the COLING 2000 Volume 1: The 18th International Conference on Computational Linguistics, Saarbrücken, Germany, 31 July 2000–4 August 2000.
54. Kupiec, J.; Pedersen, J.; Chen, F. A trainable document summarizer. In Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR ’95), New York, NY, USA, 9–13 July 1995; pp. 68–73. [\[CrossRef\]](#)
55. Ravaut, M.; Chen, H.; Zhao, R.; Qin, C.; Joty, S.; Chen, N. PromptSum: Parameter-Efficient Controllable Abstractive Summarization. *arXiv* **2023**, arXiv:2308.03117. [\[CrossRef\]](#)
56. Afzal, A.; Vladika, J.; Braun, D.; Matthes, F. Challenges in Domain-Specific Abstractive Summarization and How to Overcome Them. In Proceedings of the 15th International Conference on Agents and Artificial Intelligence, SCITEPRESS—Science and Technology Publications, Lisbon, Portugal, 22–24 February 2023; pp. 682–689. [\[CrossRef\]](#)
57. Gunturu, A.; Jadon, S.; Zhang, N.; Faraji, M.; Thundathil, J.; Ahmad, T.; Willett, W.; Suzuki, R. RealitySummary: Exploring On-Demand Mixed Reality Text Summarization and Question Answering using Large Language Models. *arXiv* **2024**, arXiv:2308.03117. [\[CrossRef\]](#)
58. Tian, Y.; Xia, F.; Song, Y. Dialogue Summarization with Mixture of Experts based on Large Language Models. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Bangkok, Thailand, 11–16 August 2024; Ku, L.W., Martins, A., Srikumar, V., Eds.; pp. 7143–7155. [\[CrossRef\]](#)
59. Brown, H.; Shokri, R. How (Un)Fair is Text Summarization? 2023. Available online: <https://openreview.net/forum?id=-UsbRIXzMG> (accessed on 30 November 2024).
60. Zhang, X.; Li, J.; Chi, P.W.; Chandrasegaran, S.; Ma, K.L. ConceptEVA: Concept-Based Interactive Exploration and Customization of Document Summaries. In Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI ’23), Hamburg, Germany, 23–28 April 2023; ACM: New York, NY, USA, 2023; pp. 1–16. [\[CrossRef\]](#)
61. Munaf, M.; Afzal, H.; Iltaf, N.; Mahmood, K. Low Resource Summarization using Pre-trained Language Models. *arXiv* **2023**, arXiv:2310.02790. [\[CrossRef\]](#)

62. Li, J.; Zhang, Y.; Karas, Z.; McMillan, C.; Leach, K.; Huang, Y. Do Machines and Humans Focus on Similar Code? Exploring Explainability of Large Language Models in Code Summarization. In Proceedings of the 32nd IEEE/ACM International Conference on Program Comprehension (ICPC '24), Lisbon, Portugal, 15–16 April 2024; ACM: New York, NY, USA, 2024; pp. 47–51. [\[CrossRef\]](#)
63. Cao, S.; Wang, L. AWESOME: GPU Memory-constrained Long Document Summarization using Memory Mechanism and Global Salient Content. *arXiv* **2023**, arXiv:2305.14806. [\[CrossRef\]](#)
64. Guo, Y.; August, T.; Leroy, G.; Cohen, T.; Wang, L.L. APPLS: Evaluating Evaluation Metrics for Plain Language Summarization. *arXiv* **2024**, arXiv:2305.14341. [\[CrossRef\]](#)

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

Hyperparameter Tuning for Domain-Specific Text Generation: A Case Study with FLAN-T5

Ms. Anukriti Kaushal
Computer Science and Engineering
Delhi Technological University
Delhi, India
anukritikaushal@dtu.ac.in

Mr. Rishabh Chauhan
Computer Science and Engineering
Delhi Technological University
Delhi, India
rishabhchauhan_23cse13@dtu.ac.in

Abstract—In this study, the legal domain’s FLAN-T5, an instruction-tuned transformer model, is used to optimize performance for domain-specific summarization. When it came to summarizing U.S. legislative texts from the BillSum dataset, FLAN-T5 originally performed poorly, falling well short of the BART-Large model despite its promising architecture. Traditional methods like pretraining on more summary data and enlarging the dataset did not produce appreciable gains. But by methodically changing two important generation parameters, *num_beams* and *generation_max_length*, significant improvement was seen in the ROUGE-1 score, which went from 18.39 to 40.85, matching BART-Large’s performance. These findings demonstrate how important hyperparameter tuning is to maximizing the potential of instruction-tuned models for domain-specific applications, especially in intricate domains like summarizing legal documents.

Index Terms—Instruction Tuning, FLAN (Fine-Tuned Language Model), XSum, BillSum, Hyperparameter

I. INTRODUCTION

In natural language processing (NLP), automatic text summarization has emerged as a crucial task that makes it possible to condense vast amounts of text into brief, insightful summaries. Bibliometric analysis of text summarisation [15] trends shows that advances in artificial intelligence continue to take forward the growth of text summarisation research, particularly with the evolution from traditional extractive approaches to cutting-edge abstractive models like BERT and GPT. Abstractive summarization, in which models create summaries by analyzing and paraphrasing the source text, has advanced significantly with the introduction of transformer-based models. The performance of these models can, however, differ greatly based on the task and dataset, which raises concerns about their generalizability and suitability for certain domains, such as summarizing legal documents.

By training on a variety of instruction–response combinations, pre-trained language models can be fine-tuned to obey natural language instructions. This technique is known as Instruction tuning. Instruction-tuned models generalize better across different tasks with less supervision than typical task-specific fine-tuning. Similar to this, domain-specific summarization entails modifying summarization models to function well in a given subject, such as financial, medical, or legal texts, which frequently contain complicated structures and specialist

vocabulary. These objectives require models that are accurate, contextually aware, and fluent in the domain.

The performance of two cutting-edge summarizing models, Flan-T5-Base [2] and BART-Large, on the BillSum legal summarization dataset is examined in this work. Their ROUGE [1] scores showed a sharp difference in the first experiments, with BART-Large significantly exceeding Flan-T5 Base (40.09 vs. 18.39). This disparity led us to look at ways to enhance Flan-T5 Base, the failing model’s performance.

Assuming that exposure to a diverse dataset may increase Flan-T5 Base’s generalization capabilities, initial steps were to try to improve its performance by fine-tuning it successively on the XSum dataset [14] and then BillSum. This strategy, though, produced no better results. Assuming that a larger dataset could aid the model in learning better representations, further investigation was done in the effects of expanding the BillSum [10] dataset. Remarkably, this did not raise the ROUGE score either. At last, focus came on to hyperparameter tuning, specifically modifying the *num_beams* and *generation_max_length* parameters. As a result of this strategy, Flan-T5 Base’s ROUGE score increased dramatically from 18.39 to 40.85.

Our results demonstrate that the performance of newer models can be highly dependent on task-specific setups and that they do not necessarily outperform older models on all datasets. This study offers insights into optimizing summarization models for domain-specific tasks and emphasizes the significance of hyperparameter adjustment [6]. The study advances the knowledge of model adaptability and optimization in NLP by showing that specific modifications can close the performance gap between models. Using the BillSum dataset, this study solely examines the legal area. The ROUGE-1 measure is the key tool used to assess model performance.

II. LITERATURE REVIEW

Recent developments in transformer-based models have shown state-of-the-art performance in summarization tasks, including T5 and BART [5]. The BART autoencoder is designed for abstractive summarization, while the T5 model, especially its FLAN-T5 variation, is instruction-tuned to enhance zero-shot learning across many NLP applications. Domain-specific summarizing is still difficult, for example, when it comes to

legal documents, when models developed on general-purpose summary datasets might not transfer well.

A. Instruction Tuned models

Instruction tuning [2] [3] is a fine-tuning technique that enhances the ability of language models to follow specific prompts and perform a wide range of NLP tasks. Instruction tuning allows models to better generalize across many domains by training them on a varied set of instruction-response pairings, as opposed to classical fine-tuning, which optimizes models for a single task. Models such as FLAN-T5, which showed notable gains in zero-shot and few-shot learning capabilities, popularized this method. Instruction tuning improves performance on tasks like question answering, summarizing, and code creation by explicitly training models on human-like instructions that better match their outputs with user expectations. Instruction tuning [4] increases flexibility, but it does not always ensure better performance in domain-specific applications because models may still need parameter tweaks and task-specific fine-tuning to achieve the best results.

B. Denoised Language models

Transformer-based architectures known as denoising language models [5] are trained to reconstruct damaged or missing text, enhancing their capacity to produce meaningful and cogent content. These models, like BART (Bidirectional and Auto-Regressive Transformer), go through a pretraining phase in which the model learns to rebuild the original sequence after parts of the input text are randomly masked, jumbled, or removed. The model may gain a profound contextual understanding thanks to this denoising aim, which makes it very useful for tasks involving abstractive summarization, text production, and translation. Denoising models use bidirectional encoding to capture relationships over an entire input sequence, in contrast to typical autoregressive models that predict tokens sequentially. They are able to produce more fluid and factually sound writing as a result. For high-stakes applications like legal and medical text summarization, thorough examination is required due to their potential disadvantage of hallucination, which causes them to produce content that is believable but factually erroneous.

C. Related Works

The potential and constraints of summarization models have been examined in a number of recent research. In order to provide a unifying method, Raffel et al. [3] created the T5 framework, which approaches all NLP tasks as text-to-text problems and requires significant resources for fine-tuning. Chung and associates. Although domain-specific performance remained uneven, [2] scaled instruction tweaking across large model families, enhancing generalization in zero-shot circumstances. By using denoising targets, Lewis et al. [5] showed how good BART is for summarizing, but they also pointed us that it is susceptible to hallucinations in delicate areas. In their survey of instruction-tuning techniques, Zhang et al. [4] found a weakness in strong domain adaptation. Hyperparameter

adjustment has a substantial impact on performance outcomes, particularly in task-specific or low-resource environments, as Liu and Wang [6] highlighted. Together, these efforts show that although huge pre-trained models have made significant progress, their practical applicability—particularly in specialized domains like legal summarization—relies substantially on contextual fine-tuning and parameter settings. By methodically assessing how hyperparameter adjustment affects FLAN-T5's performance in a domain-specific task, this study expands on those findings.

D. Model Limitations

Current summarization models have a number of drawbacks despite their improvements. Models trained on general-purpose datasets frequently struggle with specialized content, such legal or medical texts, which leads to decreased accuracy. This is a major worry due to their limited domain flexibility. The hallucination problem is another crucial issue, especially in abstractive models like GPT or BART, where generated summaries may seem fluid but contain inaccurate facts. Furthermore, a lot of models require a lot of fine-tuning, which might not be possible in settings with limited resources or for businesses with little computing power. These difficulties show that in order to make summarization systems more robust and widely applicable, more dependable domain-specific training methodologies and lightweight, parameter-efficient tuning techniques are required.

III. METHODOLOGY

This work attempts to increase the ROUGE [1] score of a summarization model by experimenting with alternative fine-tuning and setup procedures. From the huggingface library [13], two transformer-based models were chosen for comparison:

1. Model 1: FLAN-T5 Base
2. Model 2: BART-Large

The BillSum dataset [10], which includes summaries of legal publications, was used to refine both models. Model 1's ROUGE score after initial fine-tuning on BillSum was 18.39, but Model 2's ROUGE score of 40.09 was noticeably higher.

A. Improving FLAN-T5 Base

1.1 Fine-Tuning on Additional Summarization Dataset

Hypothesis: Model 1's summarizing performance may be enhanced by training it on the XSum dataset before optimizing it on BillSum.

Implementation: XSum was used to fine-tune Model 1 then BillSum was used.

Rouge Score: 25.009

Result: There was no noticeable rise in the ROUGE score.

1.2 Increasing the Dataset Size of BillSum dataset

Hypothesis: By adding more legal domain-specific instances to the BillSum dataset, Model 1's capacity to summarize legal texts would be enhanced.

Implementation: Training was done on a bigger subset of BillSum.

Rouge Score: 18.19

Result: Model 1's ROUGE score showed no discernible improvement, suggesting that dataset size by itself was not a limiting factor.

1.3 Optimization of Hyper-Parameters

Hypothesis: The quality of the summary could be enhanced by modifying the hyperparameters.

Implementation: Hyperparameters were changed for *num_beams* to 6 and *generation_max_length* to 300 in order to encourage a more thorough investigation of possible summaries.

Rouge Score: 40.85

Result: Significant improvement in Rouge Score.

B. Ensuring Reliability

The study uses Pre-trained models (FLAN-T5 Base and BART-Large) from the Hugging Face Transformers library and publically available datasets (BillSum and XSum) to increase the dependability of our findings. Consistent hardware configurations and fixed random seeds(seed = 42) were used in every trial. To guarantee stability, training and evaluation procedures were carried out four times, with performance data being noted at each run.

IV. RESULTS

Tables and graphs in this section show rouge [1] score, Training Loss and Validation Loss metrics for improvements suggested in methodology section. The graphs shown in the section are made using seaborn [9] python package. Stable and dependable model performance under repeated settings was indicated by the given results, which were consistent across epochs and displayed little volatility. Here is a brief explanation of columns given in the tables below :

Epoch - The training cycle number. Each epoch represents one full pass through the training dataset.

Training Loss - The model's loss on the training set. Lower values indicate better model fit.

Validation Loss - The model's loss on the validation set, used to assess generalization performance.

Rouge1 - Measure of overlap of unigrams (individual words) between predicted and reference summaries.

In table I, Bart-large performs well as training loss drops from 1.4 to 1.08. Validation loss drops from 1.4 to 1.3. Rouge Score increases from 39.8 to 40.0.

TABLE I

BART-LARGE ON BILLSUM WITH 8000 ROWS IN TRAIN SET AND 1000 ROWS IN TEST SET

Epoch	Training Loss	Validation Loss	Rouge1
1	1.4711	1.430512	39.8935
2	1.2866	1.375137	39.9252
3	1.1315	1.372435	39.9512
4	1.0851	1.367797	40.0978

In table II, Flan-T5 performs less than Bart-Large as training loss drops from 1.8 to 1.6. Validation loss drops from 1.7 to 1.6. Rouge Score increases from 18.2 to 18.3.

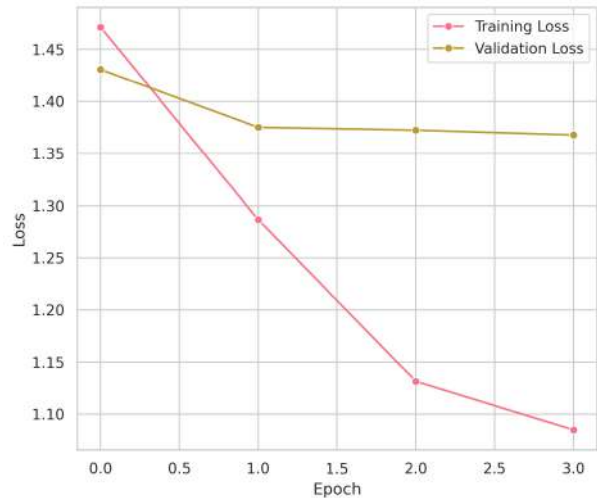


Fig. 1. Loss Graph for table I

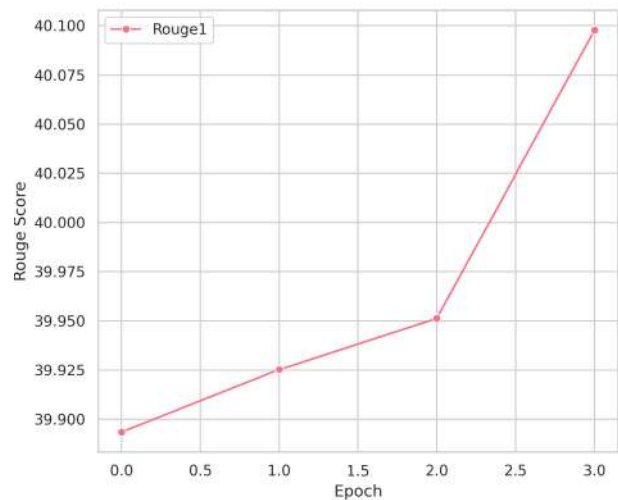


Fig. 2. Rouge Score Graph for table I

TABLE II

FLAN-T5-BASE ON BILLSUM WITH 8000 ROWS IN TRAIN SET AND 1000 ROWS IN TEST SET

Epoch	Training Loss	Validation Loss	Rouge1
1	1.8348	1.740196	18.2941
2	1.8827	1.691753	18.2926
3	1.6701	1.674098	18.4369
4	1.65	1.670359	18.3968

In table III, Flan-T5 performs slightly well as training loss drops from 1.28 to 1.27. Validation loss drops from 1.19 to 1.17. Rouge Score increases from 24.9 to 25.0.

In table IV, Flan-T5 performs almost same as in table II as

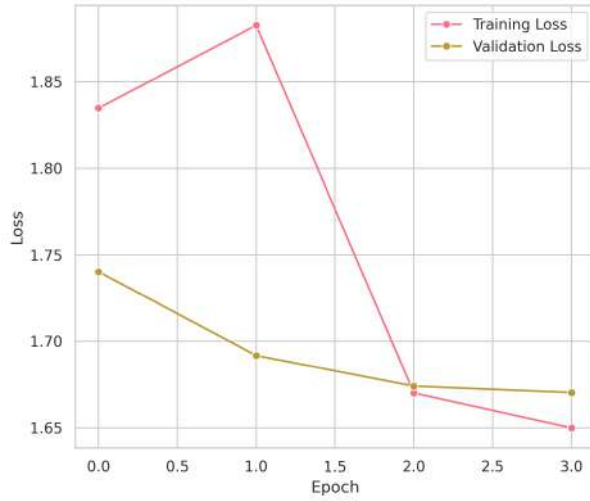


Fig. 3. Loss Graph for tableII

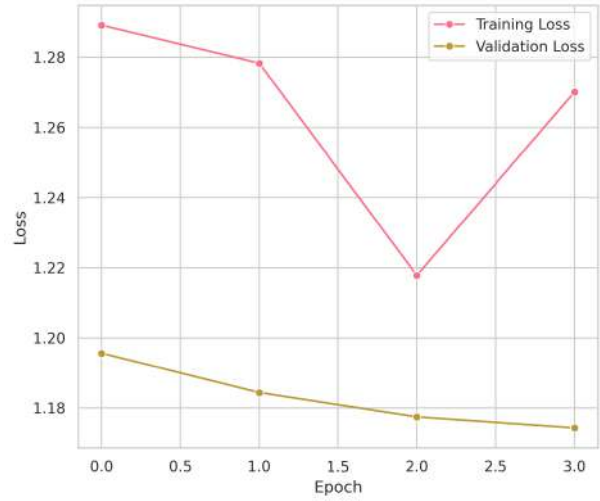


Fig. 5. Loss Graph for tableIII

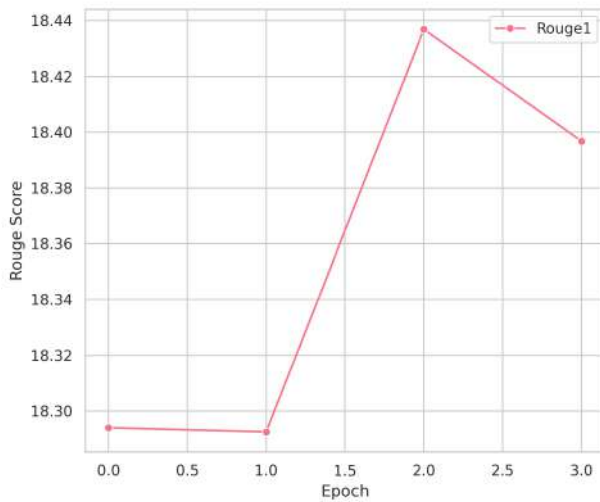


Fig. 4. Rouge Score Graph for tableII

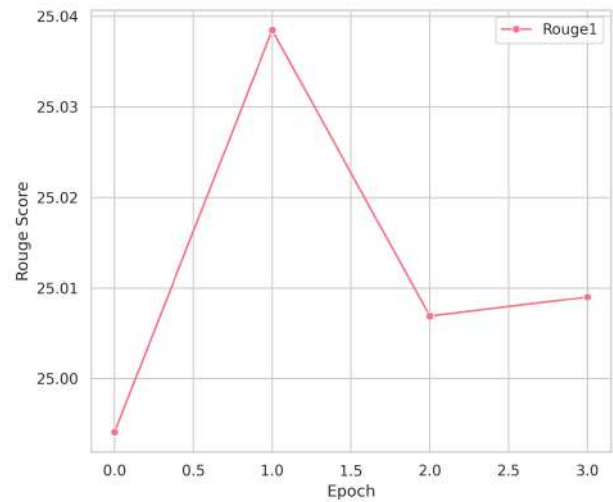


Fig. 6. Rouge Score Graph for tableIII

TABLE III
FLAN-T5-BASE ON BILLSUM, WITH 8000 ROWS IN TRAIN SET AND 1000 ROWS IN TEST SET, AFTER TRAINING ON XSUM

Epoch	Training Loss	Validation Loss	Rouge1
1	1.2891	1.195633	24.9941
2	1.2783	1.184432	25.0385
3	1.2178	1.177504	25.0069
4	1.2701	1.174383	25.009

TABLE IV
FLAN-T5-BASE ON BILLSUM WITH 18.9K ROWS IN TRAINING AND 3.27K ROWS IN TEST SET.

Epoch	Training Loss	Validation Loss	Rouge1
1	2.0935	1.870951	17.9862
2	1.9908	1.823429	18.1394
3	2.0161	1.802811	18.1797
4	1.9716	1.797551	18.1915

training loss drops from 2.0 to 1.9. Validation loss drops from 1.8 to 1.7. Rouge Score increases from 17.9 to 18.1.

In table V, Flan-T5 performs as good as Bart-Large in table I as training loss drops from 2.1 to 2.0. Validation loss drops

from 1.9 to 1.8. Rouge Score increases from 39.1 to 40.8.

These results are consistent with earlier findings by Liu and Wang [6], who emphasized how hyperparameter adjustment significantly affects model performance, especially in domain-

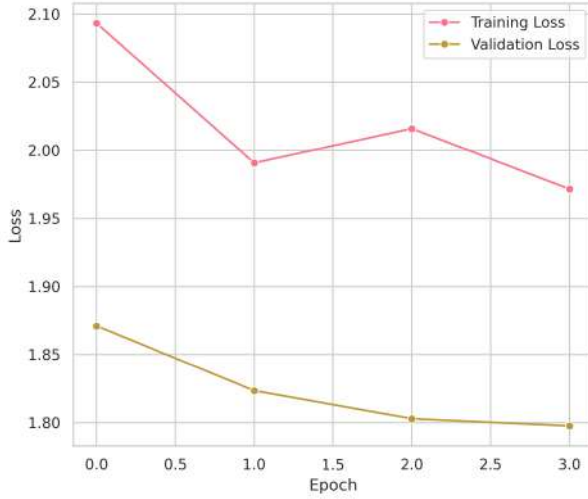


Fig. 7. Loss Graph for tableIV

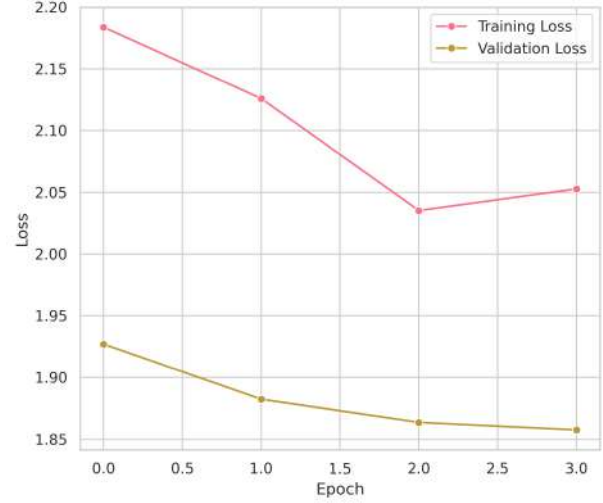


Fig. 9. Loss Graph for tableV

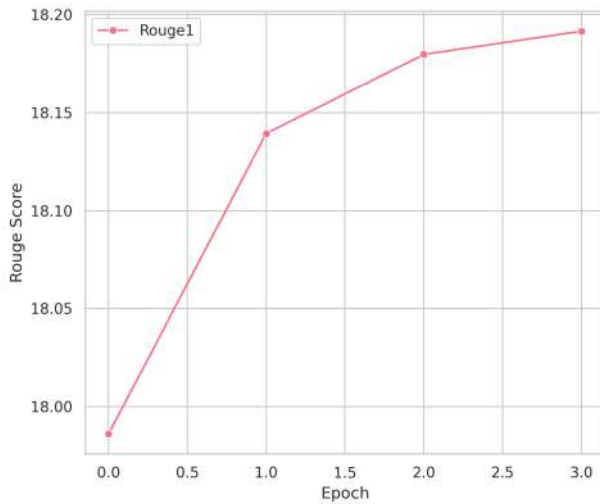


Fig. 8. Rouge Score Graph for tableIV

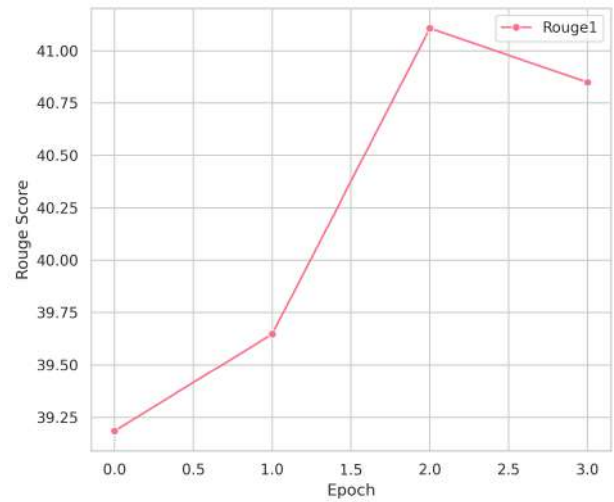


Fig. 10. Rouge Score Graph for tableV

TABLE V
FLAN-T5-BASE ON BILLSUM AFTER HYPERPARAMETER TUNING

Epoch	Training Loss	Validation Loss	Rouge1
1	2.1837	1.926986	39.1855
2	2.126	1.88247	39.6472
3	2.0353	1.863689	41.1076
4	2.0528	1.857668	40.8503

specific applications. According to research by Raffel et al. [3] and Chung et al. [2], instruction-tuned models like T5 and FLAN-T5 provide advantages in generalization across tasks, but without further modification, they frequently perform poorly in specialized areas. Our findings support this notion

by showing that fine-tuning FLAN-T5 on extra data (XSum) only slightly improved performance, confirming earlier findings that exposure to cross-domain data is not enough. The performance difference between FLAN-T5 and BART-Large was substantially reduced by targeted tuning of *num_beams* and *generation_max_length*, however, offering empirical evidence that configuration—rather than architecture or dataset size alone—is frequently the limiting factor in domain-specific summarization tasks.

V. CONCLUSION

In some cases, older models continue to outperform newer ones, as observed in this study where FLAN-T5

initially underperformed on legal summarization tasks despite being instruction-tuned. Attempts to improve its performance through additional dataset fine-tuning and dataset expansion did not yield significant gains, indicating that the volume of training data was not the primary constraint. However, a substantial improvement in ROUGE scores was achieved by adjusting key hyperparameters such as *generation_max_length* and *num_beams*, demonstrating the critical role of hyperparameter tuning in enhancing summarization quality. This adjustment significantly narrowed the performance gap between FLAN-T5 Base and BART-Large. Minimal increases were obtained from early attempts to increase performance by more pretraining and dataset expansion, suggesting that data was not the only limiting factor. Targeted hyperparameter adjustment, however, produced notable gains, increasing the ROUGE-1 score from 18.39 to 40.85. This performance was comparable to that of the more robust baseline, BART-Large. Crucially, this study offers new perspectives on how to optimize instruction-tuned, lightweight models in domain-specific, low-resource settings. Our findings emphasize the importance of model-specific hyperparameter tuning for domain-specific summarization, such as legal text. Our approach's main advantage is that it uses publicly accessible models and datasets to generate notable performance gains without requiring extensive retraining. This makes it a workable and repeatable solution for real-world, low-resource applications.

VI. LIMITATIONS

Despite the fact that this study effectively illustrates how hyperparameter adjustment can enhance summarization performance, a number of restrictions still exist. First off, BillSum, the sole legal-domain dataset used for the evaluation, might not accurately reflect the variety and complexity of legal documents that are encountered in everyday situations. Therefore, the findings' generalizability to other legal or domain-specific corpora (such as technical, financial, or medical literature) has not been explored.

Second, even though hyperparameter tuning significantly improved FLAN-T5 Base, it still required manual experimentation with parameters like *num_beams* and *generation_max_length*, which might not be the best for all workloads or datasets. Performance benefits might have been missed since the tuning procedure was heuristic and lacked systematic hyperparameter optimization methods like grid search or Bayesian optimization.

Furthermore, ROUGE ratings were the only evaluation metric used in the study. Despite its widespread use, ROUGE falls short in capturing the readability, factual accuracy, and semantic fidelity of generated summaries, particularly in delicate fields like law. A more thorough analysis that makes use of human evaluations or extra measures (such as BERTScore [11] or factual consistency checks) would offer a more full picture of model performance.

Lastly, the study did not investigate retrieval-augmented summarization strategies or parameter-efficient fine-tuning

techniques (such as LoRA [12], adapters, or quick tuning), which could improve the performance of underperforming models even more without consuming a lot of computational power.

Cross-validation across different datasets and hardware configurations could further enhance external reliability and generalization assurance, even if the trials were reproducible within our testing environment. By collecting a wider variety of datasets, using more methodical tuning techniques, incorporating a range of assessment measures, and investigating more recent tuning and design solutions for robust domain adaptation, future work should overcome these limitations.

REFERENCES

- [1] ROUGE: A Package for Automatic Evaluation of Summaries(<https://aclanthology.org/W04-1013/>) (Lin, 2004)
- [2] Chung, Hyung Won, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li et al. "Scaling instruction-finetuned language models." *Journal of Machine Learning Research* 25, no. 70 (2024): 1-53.
- [3] Raffel, Colin, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. "Exploring the limits of transfer learning with a unified text-to-text transformer." *Journal of machine learning research* 21, no. 140 (2020): 1-67.
- [4] Zhang, Shengyu, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li et al. "Instruction tuning for large language models: A survey." *arXiv preprint arXiv:2308.10792* (2023).
- [5] Lewis, Mike, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov and Luke Zettlemoyer. "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension." *Annual Meeting of the Association for Computational Linguistics* (2019).
- [6] Liu, Xueqing, and Chi Wang. "An empirical study on hyperparameter optimization for fine-tuning pre-trained language models." *arXiv preprint arXiv:2106.09204* (2021).
- [7] Ülker M, Özer AB (2024) The Bart-based Model for Scientific Articles Summarization. *JUCS - Journal of Universal Computer Science* 30(13): 1807-1828. <https://doi.org/10.3897/jucs.115121>.
- [8] Abdul Ghafoor Etemad, Ali Imam Abidi, and Megha Chhabra. Fine-Tuned T5 for Abstractive Summarization [J]. *Int J Performability Eng*, 2021, 17(10): 900-906.
- [9] Waskom, M. L., (2021). seaborn: statistical data visualization. *Journal of Open Source Software*, 6(60), 3021, <https://doi.org/10.21105/joss.03021>.
- [10] Anastassia Kornilova and Vladimir Eidelman. 2019. BillSum: A Corpus for Automatic Summarization of US Legislation. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 48–56, Hong Kong, China. Association for Computational Linguistics.
- [11] Zhang, Tianyi, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. "Bertscore: Evaluating text generation with bert." *arXiv preprint arXiv:1904.09675* (2019).
- [12] Hu, Edward J., Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. "Lora: Low-rank adaptation of large language models." *ICLR* 1, no. 2 (2022): 3.
- [13] Wolf, Thomas, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac et al. "Huggingface's transformers: State-of-the-art natural language processing." *arXiv preprint arXiv:1910.03771* (2019).
- [14] Narayan, Shashi, Shay B. Cohen, and Mirella Lapata. "Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization." *arXiv preprint arXiv:1808.08745* (2018).
- [15] Kaushal, Anukriti, Chia-Chen Lin, Rishabh Chauhan, and Rajeev Kumar. 2024. "Charting the Growth of Text Summarisation: A Data-Driven Exploration of Research Trends and Technological Advancements" *Applied Sciences* 14, no. 23: 11462. <https://doi.org/10.3390/app142311462>