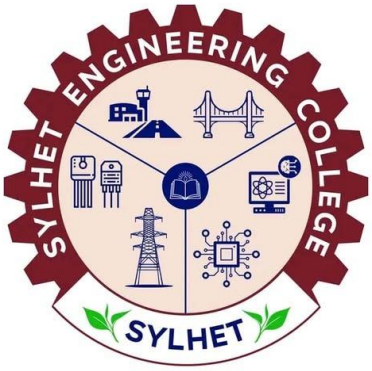
SYLHET ENGINEERING COLLEGE

Affiliated with Shahjalal University of Science and Technology

Department of Computer Science & Engineering

CSE 800



Simplifying Legal Texts: Bangla Deed Summarization Approach

Submitted by

Sifat Samin Sarker
Reg. No.: 2019331550

Eftiar Alam Patwary
Reg. No.: 2019331561

Department of Computer Science & Engineering

Supervisor

Md. Abu Naser Mojumder
Assistant Professor and
Head of CSE Department
Department of Computer Science & Engineering
Sylhet Engineering College

Recommendation Letter from Thesis Supervisor

The thesis entitled “Simplifying Legal Texts: Bangla Deed Summarization Approach” submitted by the students

1. Sifat Samin Sarker
2. Eftiar Alam Patwary

is a record of research work carried out under my supervision and I, hereby, approve that the report is submitted in partial fulfilment of the requirement for the award of their Bachelor’s Degree.

– –

Signature of the Supervisor

Md. Abu Naser Mojumder

Assistant Professor and

Head of CSE Department

Department of Computer Science & Engineering

Sylhet Engineering College

Acknowledgement

We would like to express our heartfelt gratitude to the **Department of Computer Science and Engineering, Sylhet Engineering College** for providing us with the resources and academic environment necessary to carry out this research.

We are especially thankful to our honorable supervisor **Md. Abu Naser Mojumder** for his invaluable guidance, continuous support and insightful feedback throughout every stage of our thesis. His encouragement and suggestions have been instrumental in shaping the direction of our work.

We also extend our appreciation to all the respected fellow researchers for their previous contributions which enriched the depth of our study.

Simplifying Legal Texts: Bangla Deed Summarization Approach

by

Sifat Samin Sarker, Eftiar Alam Patwary

Submitted to the Department of Computer Science & Engineering, in partial fulfillment of the requirements for the degree of
Bachelor of Science in Computer Science & Engineering

Abstract

Land deeds are vital legal documents establishing the property ownership of any individual. The deeds consist of complex legal language and non standard formats. Bangla land deeds are specially difficult to interpret. This challenge is magnified in Bangladesh where land deeds are often handwritten, poorly scanned or contain difficult to understand legal language. While Natural Language Processing (NLP) and Large Language Models (LLMs) have advanced legal document understanding overall, Bangla still remains a low resource language specially in the legal domain. This research offers a pipeline for abstractively summarizing Bangla land deeds by combining fine-tuned LLMs with optical character recognition (OCR). Using Pytesseract OCR, the pipeline first extracts Bangla text from scanned deed images followed by manual correction and text cleaning. Then using a custom dataset of land deeds and hand written summaries, the two transformer based models mT5 and BanglaT5 are fine-tuned for summarization. The models were evaluated using surface level metrics (CER, WER, BLEU), semantic alignment (BERTScore) and structural accuracy (Exact Match). Although both models peforrmed well, mT5 excelled in fluency and surface level accuracy while BanglaT5 preserved legal phrasing and was more precise. This research presents a workable approach to ensure all Bangla land deeds are easily accessible while contributing to the field of legal informatics in low resource languages and providing a solid foundation for future advancements.

Table Of Content

Content Page

Chapter 1: Introduction ---------- 1 - 7

1.1 Background ----------2

1.2 Problem Statement----------2

1.3 Motivation ----------3

1.4 Research Area----------3

1.5 Research Aim ----------4

1.6 Research Objectives ----------4

1.7 Scope and Limitations ---------- 5

1.8 Thesis Organization----------6

Chapter 2: Background Study---------- 8 - 11

2.1 The Importance of Legal Document Understanding----------8

2.2 Evolution of Text Summarization Techniques---------- 8

2.3 Advances in Bangla Text Summarization---------- 9

2.4 The Challenge of Legal Document Summarization---------- 9

2.5 OCR Technologies in Document Digitization ---------- 10

2.6 The Landscape of Land Deeds in Bangladesh ---------- 10

2.7 Synthesis and Knowledge Gap---------- 11

Chapter 3: Literature Review---------- 12 - 18

3.1 Evolution of Text Summarization ----------12

3.2 Bangla Text Summarization ---------- 13

3.3 Legal Document Summarization---------- 13

3.4 OCR Advances—---------- 14

3.5 Bangla Text OCR---------- 15

3.6 Land Deeds and Their Summarization ---------- 16

3.7 Related Works----------16

3.8 Research Gaps ---------- 17

Chapter 4: Dataset ---------- 19 - 26

4.1 Dataset Overview---------- 19

4.2 Data Collection ---------- 19

4.3 Dataset Samples ---------- 20

4.4 Preprocessing and Cleaning---------- 25

4.5 Annotation Process---------- 26

4.6 Ethical Considerations ----------26

Chapter 5: Methodology ---------- 27 - 33

5.1 Data Overview ---------- 28

5.2 OCR Preprocessing ----------28

5.3 Text Extraction and Summary Modeling---------- 28

5.4 Model Selection and Adaptation---------- 28

5.5 Deed Summarization ----------29

5.6 Evaluation ---------- 30

5.7 Workstation Configuration----------31

5.8 Experimental Setup and tools used---------- 32

Chapter 6: Results and Discussion ---------- 34 - 43

6.1 Evaluation Metrics and Strategy ---------- 34

6.2 Quantitative Results ---------- 35

6.3 Analysis and Interpretation ---------- 36

6.4 Model Behavior Observations ---------- 37

6.5 Visual Analysis ---------- 37

6.6 Summary of Comparative Performance ---------- 42

6.7 Implications for Legal Text in Bangla----------42

6.8 Model Determining Table ---------- 43

Chapter 7: Conclusion and Future Work ---------- 44 - 46

7.1 Conclusion ----------44

7.2 Future Work ----------44

References ---------- 47 - 51

List of Figures

Figure 1.1: Research Area Diagram ----------4

Figure 4.1: Text Length Distribution for Deed Dataset ---------- 20

Figure 4.2: Sample Deed Page 1 ---------- 20

Figure 4.3: Extracted text or Input text ---------- 21

Figure 4.4: Manually Summarized Text---------- 21

Figure 4.5: Sample Deed Page 2 ---------- 22

Figure 4.6: Sample Deed Page 3 ---------- 22

Figure 4.7: Sample Deed Page 4 ---------- 23

Figure 4.8: Sample Deed Page 5 ---------- 23

Figure 4.9: Extracted Text Sample 1 ----------23

Figure 4.10: Extracted text sample 2 ---------- 24

Figure 4.11: Extracted text sample 3 ---------- 24

Figure 4.12: Extracted text sample 4 ---------- 24

Figure 4.13: Summarized text sample 1 ---------- 24

Figure 4.14: Summarized text sample 2 ---------- 24

Figure 4.15: Summarized text sample 3 ---------- 25

Figure 4.16: Summarized text sample 4 ---------- 25

Figure 4.17: Word Cloud for input_text and target_text(summary) in Dataset---------- 26

Figure 5.1: Methodology Overview—----------27

Figure 6.1: Model Performance Comparison Across Metrics –---------- 35

Figure 6.2: Training vs Validation Loss Curve (banglaT5 Model) ---------- 38

Figure 6.3: Training vs Validation Loss Curve (mT5 Model) ---------- 39

Figure 6.4: Training vs Validation Loss: BanglaT5 vs mT5---------- 41

List of Tables

Table 6.1: Quantitative Results of mT5 and BanglaT5 ----------35

Table 6.2: Comparative Visualization and Interpretation of mT5 and BanglaT5--------40

Table 6.3: Comparative Performance of mT5 and BanglaT5 ----------42

Table 6.4: Model Determining Table —----------43

Chapter 1

Introduction

Legal documents serve as the foundation for governance, property ownership and civic administration in any society. Among them, land deeds are some of the most essential in establishing proof of ownership and containing details about legal agreements, boundaries, previous owners and more. However, these documents are often lengthy, written in legalese and difficult to navigate for the average citizen. This issue becomes even more prominent in Bangladesh, where land deeds are usually written in Bangla and often follow inconsistent formats with outdated or overly formal language. For many individuals, from rural farmers to urban landowners, simply understanding what their deed states can be a daunting task.

In recent years, Natural Language Processing (NLP) and machine learning [4], especially through Large Language Models (LLMs), have made it increasingly feasible to automate the process of understanding and generating human language [1][2]. When applied to the legal domain, these technologies can potentially simplify access to legal knowledge by simplifying documents that would otherwise require expert guidance. Despite the rapid progress in NLP for English and other major languages, Bangla remains a low resource language with limited available datasets, tools and pre-trained models particularly in the context of the legal domain [3].

Our research aims to bridge this gap by focusing on Bangla land deeds that are notoriously difficult to understand and largely untouched by modern NLP tools. A pipeline is proposed by us that begins with optical character recognition (OCR) to extract Bangla text from scanned deed images and then applies abstractive summarization using fine-tuned LLMs to produce easily readable summaries [5][9][10]. This thesis presents our journey towards building and evaluating such a system, contributing to the broader goals of legal technology and effortless access to information in under-represented languages.

1.1 Background

Summarization is an essential tool for navigating complex legal documents efficiently. It is really helpful for researchers, registry clerks and deed writers who need to quickly access relevant information buried within convoluted content. More critically, summarization can empower non-expert users who may not know what specific information they need from a deed[5]. A high quality summary provides them with a clear understanding of what the document is about, which is often enough to make important decisions or initiate further action [11][12].

In the case of Bangla land deeds, this is a significant issue. These documents are often difficult to comprehend for the average person due to their dense language and outdated phrasing. Despite their importance in everyday life, from property transfers to legal disputes, Bangla deeds have not been explored in the context of automatic summarization at a large scale. Our approach first uses pytesseract OCR to extract raw Bangla text from scanned deed documents [8]. This is followed by abstractive summarization, in which a generative model creates a concise and meaningful summary that captures the essence of the document [13].

1.2 Problem Statement

The structure and language of Bangla land deeds pose significant challenges for the general public. These documents are long, contain complex legal expressions and are often written in a way that is difficult to understand even for educated readers. Manual summarization of these deeds by legal professionals or intermediaries is not only reliant on a lot of time but also subject to inconsistencies in interpretation, leading to misinformation or misunderstanding.

Given that no standard or technological method currently exists to simplify and summarize Bangla land deeds in an abstractive way, there is a clear and urgent need to work towards a technological solution that can process these documents and produce readable summaries accessible to everyone.

1.3 Motivation

The inspiration for this research is rooted in our personal experiences. While going through our own land related documents and legal papers, some obstacles were faced like locating key information or understanding what certain clauses meant. The dense legal language and inconsistent formatting made the documents hard to understand. This challenge led us to question how difficult understanding these documents must be for everyone, especially those who may not have the time, education or resources to seek professional legal help.

This realization sparked the idea of applying NLP techniques to automatically summarize legal documents that works towards making legal content more approachable. The field of scope was narrowed down to land deeds given their significance in daily life and the high frequency with which people interact with them.

One of the biggest challenges faced was in data collection. Because land deeds are private and sensitive documents, people are often unwilling to share them. Even obtaining them from government sources like Land Registry Offices involves bureaucratic hurdles. The challenges were solved by collecting land deeds from the Land Office through connections and personal samples. Maintaining a strict standard of data ethics and privacy has been a top concern while handing these documents.

1.4 Research Area

This research lies at the intersection of Natural Language Processing (NLP), Bangla deed document understanding and legal informatics, with a special focus on abstractive summarization[14]. While OCR plays an important supporting role, our main emphasis is on using generative models (LLMs) for abstractive text summarization [9][10]. Specifically, exploring how such models can be fine-tuned to understand and summarize complex, unstructured legal documents written in Bangla.

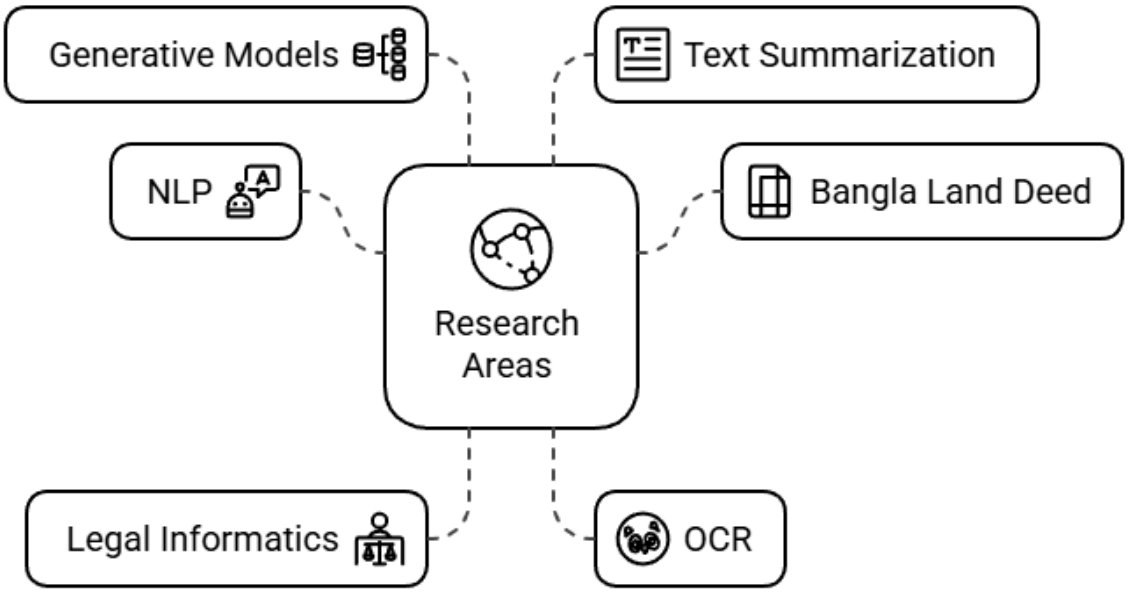


Figure 1.1: Research Area

1.5 Research Aim

The main goal of this research is to establish a complete pipeline that can extract text from images of Bangla land deeds using OCR and generate an abstractive summary, enabling users to easily understand the key content of the deed.

1.6 Research Objectives

To fulfill this aim, a number of key objectives that guide the development, evaluation and refinement of summarization approach were set. These objectives include both technical goals (such as OCR and model fine-tuning) and analytical goals (such as model comparison and performance optimization). Through these objectives, the intention is not only to work towards establishing a functional pipeline but also to contribute new insights into summarization methods for Bangla legal documents.

The objectives are as follows:

- To explore the field of Bangla text summarization, specifically applied to land deeds.
- To build an OCR pipeline that can accurately extract Bangla text from deed images.
- To fine-tune large language models (LLMs) for generating coherent and accurate summaries of Bangla land deeds.
- To evaluate and compare the performance of different fine-tuned LLMs on our custom dataset.
- To analyze the effect of hyperparameter tuning on the performance of LLM's on our custom dataset.
- To achieve the best possible summarization results based on quantitative and qualitative evaluation metrics.
- Understanding the legal parameters of land deeds from an expert in this field to make our research more effective.

1.7 Scope and Limitations

Scope

This research focuses on the abstractive summarization of Bangla land deeds using a combination of OCR and fine-tuned Large Language Models (LLMs). The key areas covered include:

- Developing a pipeline that converts scanned Bangla deed images into digitized text using OCR.
- Creating and curating a dataset of Bangla land deeds for training and evaluation.
- Fine-tuning pre-trained transformer models to generate concise, readable summaries of the deeds.
- Evaluating the performance of summarization using both quantitative metrics and qualitative analysis.
- Exploring the impact of OCR quality on summarization accuracy.

- Addressing specific challenges related to low resource Bangla NLP and domain specific text in the legal field.

This study aims to contribute to the broader field of Bangla legal document processing, with a special focus on accessibility and comprehensibility for non-experts.

Limitations

While the proposed approach aims to be comprehensive, there are several limitations to this study:

- **Dataset Availability**: Due to privacy concerns and limited public access, the dataset is relatively small and may not cover the full diversity of land deeds across Bangladesh.
- **OCR Accuracy**: OCR errors from poor quality scans can impact summarization quality and complete accuracy is beyond the scope of this thesis.
- **Model Generalization**: The fine-tuned models are trained on a specific dataset and may not generalize well to unseen deed formats and legal styles.
- **Language Scope**: The study is limited to Bangla language deeds. Multi-language deeds or translated documents are not considered.
- **Computational Constraints**: Due to hardware limitations, very large models (like GPT-4 or LLaMA-3) or exhaustive hyperparameter tuning were not feasible during experimentation. Taking a large token size resulted in a vast amount of training time which became infeasible at one point.

1.8 Thesis Organization

This thesis is structured into seven Chapters:

Chapter 1: Introduction
Outlines the research motivation, objectives, scope and challenges related to the summarization of Bangla land deeds. It presents the context of legal document accessibility issues in Bangladesh and introduces the proposed solution.

Chapter 2: Background Study

Reviews the theoretical and technical foundations of OCR, text summarization, Bangla NLP and legal document processing. It highlights the complexities of Bangla deeds and identifies existing research gaps.

Chapter 3: Literature Review

Surveys existing works in OCR, Bangla summarization, legal NLP and LLM-based text generation. This Chapter contextualizes our contributions and elaborates on the novelty of the approach.

Chapter 4: Dataset

Goes through the custom curated dataset with an overview, the collection process, preprocessing and cleaning, annotation process and the ethical considerations to be made throughout the dataset maintainance.

Chapter 5: Methodology

Describes the complete system pipeline, from data collection and preprocessing to model fine-tuning and evaluation. It provides detailed insight into dataset preparation, model selection, training procedures and evaluation strategies.

Chapter 6: Results and Discussion

Presents quantitative and qualitative performance analysis of the fine-tuned models. It discusses training trends, metric comparisons and model behavior under different input conditions.

Chapter 7: Conclusion and Future Work

Summarizes key findings, draws conclusions and outlines future directions to improve model robustness, dataset scalability and real-world deployment, especially for legal accessibility in Bangla.

Chapter 2

Background Study

Understanding the legal landscape of Bangla land deeds requires a deep dive into multiple disciplines like legal informatics, natural language processing and document digitization. This Chapter explores the foundations that underpin our approach to summarizing Bangla legal documents. Beginning with the importance of legal text comprehension, then examining the evolution of summarization techniques, OCR technologies and the unique characteristics of Bangla land deeds. Together, these chapters reveal the challenges and research gaps that our work seeks to address.

2.1 The Importance of Legal Document Understanding

Legal documents including land deeds constitute the backbone of property rights, governance and dispute resolution in every society. These documents are complex by design, filled with technical jargon, dense in structure and often obsolete in phrasing. In Bangladesh, property ownership relies heavily on Bangla land deeds. The challenge of understanding such documents is even more acute due to their non standard formats and linguistic complexity [15]. Document understanding or basically summarization, is therefore vital in reducing the cognitive and operational burden on the general public and professionals who need to interpret these documents regularly [16].

2.2 Evolution of Text Summarization Techniques

2.2.1 Extractive Summarization

The origin of text summarization lies in extractive techniques, where the goal is to select the most information packed sentences or phrases verbatim from the source text. Early algorithms relied on statistical heuristics such as term frequency, sentence position and cue phrases. Later, graph based methods like TextRank represented relationships between sentences, enhancing summary relevance and coherence. [17] While effective for short

documents or news articles, extractive methods often fail to provide abstractions suited to user needs, especially for complex legal text. [18]

2.2.2 Abstractive Summarization

Abstractive summarization marks a significant leap, as it seeks to generate new sentences that capture the core meaning of the original text, paraphrasing wherever needed. Neural approaches, especially encoder-decoder architectures and transformers like BART and T5 have shown remarkable capacity for text generation and paraphrasing. [10][19] The quality and fluency of these summaries typically surpass extractive ones, especially in domains demanding contextual understanding such as legal deed understanding. [20]

2.3 Advances in Bangla Text Summarization

Bangla is among the world's most widely spoken languages, yet it is not well resourced for NLP, primarily due to the scarcity of annotated corpora and specialized tools[21]. Early efforts concentrated on extractive summarization using rule based or statistical methods tailored to Bangla's morphology and syntax.[22] Recent studies have explored pre-trained transformer models and even attempted transfer learning from multilingual models like BERT or mT5 to Bangla, but the bottleneck remains domain specific annotated data. [23] [24] Notably, Bangla abstractive summarization is largely uncharted with only a handful of studies attempting neural text generation in this language and generally focusing on domains like news rather than law.

2.4 The Challenge of Legal Document Summarization

2.4.1 Complexity and Structure

Legal documents possess unique structural and linguistic challenges including long sentences, intricate argumentation and strict formalism. This hinders both manual and automated interpretation.

2.4.2 Large Language Models for Legal Texts

With the emergence of LLMs, legal document summarization has entered a new era. [16] Fine-tuning LLMs on domain specific corpora such as law reports or legal contracts has markedly improved the quality of summary generation for languages like English.[25] Techniques like Retrieval Augmented Generation (RAG) and parameter efficient tuning (e.g., LoRA, QLoRA) facilitate adaptation to narrow domains while minimizing computational overhead. [26] However, Bangla has seen little advancement in legal specific LLM adaptation due to the lack of annotated legal datasets and resources.

2.5 OCR Technologies in Document Digitization

Optical Character Recognition (OCR) forms the foundation for any digital workflow involving all types of physical documents.[6] For Bangla language, recognized OCR engines such as Tesseract have improved script coverage but challenges remain due to diverse handwriting styles, document aging and degraded scans.[27] Recent advancements in deep learning like convolutional neural networks and attention based models have raised Bangla OCR accuracy. [28] This was propelled partly by the release of larger annotated datasets covering a wide range of Bangla sources. Nonetheless, OCR errors remain a major source of noise for downstream tasks like legal document summarization where precision is critical. [29]

2.6 The Landscape of Land Deeds in Bangladesh

2.6.1 Structure and Language

Bangla land deeds are essential for establishing ownership, property boundaries and transactional history. However, they show little standardization in format or structure. These are often handwritten or printed and use formal or archaic Bangla legal language. Understanding their contents can be really difficult for ordinary users like rural landowners, agricultural laborers or even educated urban dwellers, emphasizing the need for assistance in interpretation. [15]

2.6.2 Existing Automation Efforts

Most technological interventions in Bangladeshi land documentation have revolved around digitization and record keeping (e.g., land registration systems). [15] Attempts to automate deed analysis have been rare and typically focused on metadata extraction rather than semantic summarization. There are very few publicly available systems that combine OCR and NLP for end-to-end abstractive summarization of Bangla land deeds. [28] [30]

2.7 Synthesis and Knowledge Gap

Despite significant progress in text summarization, the intersection of legal NLP and OCR for low resource languages like Bangla, especially in the domain of land deeds, remains severely underexplored.

Key gaps include:

- Insufficient annotated datasets for Bangla legal document summarization, restricting model training and evaluation.
- Lack of OCR integration scope with domain adapted LLMs capable of producing readable summaries from real Bangla land deed images.
- Little research on the error propagation characteristics of OCR output as input to downstream summarization for legal documents in Bangla.
- Absence of domain specific evaluation metrics or human-in-the-loop validation processes tailored to the legal information needs of Bangladeshi users.
- By addressing these knowledge and technological gaps, targeted NLP research on Bangla land deeds can simplify access to property related legal information, reduce misinterpretation and enable more transparent land governance.

Chapter 3

Literature Review

Text summarization is a fundamental task in Natural Language Processing (NLP) that generates concise representations of extensive documents, enabling faster information access while preserving essential meaning. Summarization techniques are broadly classified into extractive methods, which select major parts directly from the source and abstractive methods that produce new, paraphrased content while retaining core semantics. This Chapter reviews the evolution of summarization techniques, focusing on their developments with respect to Bangla language processing, legal document handling and OCR integration and concluding with related works and research gaps.

3.1 Evolution of Text Summarization

Research in automatic text summarization has transitioned from early statistical approaches to state of the art Large Language Model (LLM) based systems. Initially, extractive summarization exploited lexical features like word frequency, position and TF-IDF for sentence importance [31]. Latent Semantic Analysis (LSA) and graph based ranking algorithms such as TextRank further refined extractive capabilities by modeling semantic relations among sentences. As labeled datasets grew and computational power advanced, supervised machine learning models (e.g., SVMs, Random Forests) were applied to feature based summarization tasks.

Deep learning revolutionized the field by introducing encoder-decoder seq2seq models, enabling abstractive summarization that generates paraphrased summaries through neural language modeling. Variants employing attention mechanisms and transformers improved contextual understanding and fluency. Hybrid methods combining extractive and abstractive elements emerged to leverage extractive precision and abstractive creativity, enhancing summary informativeness and coherence [32].

The recent influx of large pre-trained language models such as OpenAI's GPT series (GPT-3, GPT-4), Google's T5 and mT5 and Facebook's BART demonstrated remarkable generalization for summarization across domains and languages due to their extensive

knowledge and contextual learning capabilities [33][2]. These transformer based LLMs further enhance zero-shot and few-shot summarization, reducing dependence on task specific training data.

3.2 Bangla Text Summarization

Despite extensive NLP research globally, Bangla language summarization lags behind due to data scarcity and script complexity. Initial efforts are concentrated on unsupervised extractive methodologies such as TextRank, which utilize graph based sentence ranking techniques suited for Bangla's morphological structure [21]. Fuzzy clustering combined with aggregated scoring has also been used to rank and select important sentences in Bangla summaries, handling syntactic and lexical particularities.

In recent years, transformer based models and pre-trained multilingual LLMs such as mT5 and BanglaBERT have been fine-tuned on Bangla summarization datasets, leading to notable gains. Additional developments leverage instruction tuning and domain adaptation for better alignment with Bangla syntax and semantics [20]. Evaluation on Bangla news corpora and government reports has shown that transformer models, especially when fine-tuned or combined with human-in-the-loop approaches, can produce fluent and semantically rich Bangla summaries even in zero-shot or few-shot settings [34].

However, the availability of large scale, high quality annotated datasets remains limited, posing challenges for further performance improvements.

3.3 Legal Document Summarization

Legal documents, including contracts, judgments and deeds, often contain complex terminology, lengthy narratives and structured arguments, which complicate automated summarization. Classical summarization attempts relied on extractive methods preserving key legal clauses, but these methods struggle with coherence and integration of domain specific knowledge.

Recent research employs abstractive summarization using neural models, fine-tuned on legal corpora, to generate more coherent and context aware summaries. The creation of domain

specific models such as LEGAL BERT which is a BERT variant pretrained on legal texts and integration of argument mining have yielded promising results in understanding and summarizing legal language [25]. Hybrid frameworks combining rule based extraction with neural generation have also been proposed to balance factual accuracy and readability.

However, challenges remain around domain adaptation, limited labeled data and evaluation complexities due to legalese nuance [35]. Advances in fine-tuning LLMs with legal knowledge and specialized datasets continue to improve summarization quality in this domain.

3.4 OCR Advances

Optical Character Recognition (OCR) technology plays a critical role in digitizing analog legal documents, enabling computational processing including summarization. Recent advances in OCR leverage convolutional neural networks (CNNs) and transformer architectures, which significantly improve recognition accuracy for both printed and handwritten texts, especially in complex layouts such as multi column legal forms and registries [7]. Innovations including adaptive binarization, layout analysis and text spotting models have made OCR more robust against poor document quality, complex fonts and noisy backgrounds.

OCR has transformed many governments and institutional legal workflows by automating data extraction from scanned contracts, court rulings and property deeds. High performing OCR systems facilitate faster data retrieval, support compliance monitoring and contribute to searchable digital repositories, reducing manual labor and error. Despite these advances, recognizing degraded and cursive handwriting common in legacy legal archives remains a significant challenge. Nevertheless, challenges remain for recognizing cursive or degraded handwritten documents common in legacy legal archives.

A notable open source OCR tool gaining extraction is pytesseract, a Python wrapper for the widely used Tesseract OCR engine. Pytesseract enables easy integration of OCR into legal document processing workflows, providing capabilities to extract text from scanned images or PDFs with relatively high accuracy for printed text. It supports multilingual OCR,

including Bangla and is often used in research and development projects for preliminary text extraction before further NLP processing [8][6]. However, for handwritten or complex documents, pytesseract alone may require enhancement via preprocessing or combined AI models.

3.5 Bangla Text OCR

The digitization of Bangla scripts via OCR introduces significant challenges due to Bangla's intricate ligature system, varying handwriting styles and document deterioration over time. Unlike Latin scripts, Bangla characters combine connected strokes and modifiers making segmentation and recognition difficult.

Recent datasets such as the Gold Standard Bangla OCR corpus [6] provide extensive multi author annotated samples facilitating supervised learning approaches. Neural OCR systems now incorporate convolutional recurrent architectures, attention mechanisms and lexicon driven post processing to tackle Bangla specific difficulties [28]. Additionally, specialized techniques address page segmentation issues, skew correction and non uniform font styles [6].

Pytesseract is often employed as a baseline OCR tool for Bangla printed text due to its support for the script and ease of use. However, its performance on handwritten or complex documents is limited and usually supplemented with domain specific custom models or preprocessing pipelines [8]. Achieving high accuracy on historic or degraded Bangla legal documents remains an open problem due to data scarcity and handwriting variance.

Despite these advances, achieving high accuracy for handwritten and historic Bangla legal documents remains an ongoing research challenge exacerbated by limited high quality real world data.

3.6 Land Deeds and Their Summarization

Land deeds are essential for establishing legal property rights but are often handwritten, inconsistent in format and incorporate domain specific jargon. In Bangladesh, deeds predominantly use Bangla script with legal terminology, compounding recognition and summarization difficulties [17]. The lack of standardized templates further complicates automated text processing.

Existing work tends to focus on metadata extraction or digitization efforts rather than semantic summarization of deed contents. Extractive heuristics have been applied to highlight important clauses but suffer from limited scalability and contextual understanding [28][27]. Comprehensive end-to-end systems that integrate OCR with abstractive summarization tailored for Bangla deeds remain absent in public research.

The few efforts addressing the semantic understanding of Bangla land deeds are generally isolated and suffer from data paucity and lack of legal domain adaptation, revealing a significant research gap.

3.7 Related Works

Comparative studies have advanced understanding of summarization across languages, domains and tasks, while novel evaluation techniques aim to better capture summary quality and fidelity.

Cross lingual summarization approaches combine translation and summarization pipelines or adopt multilingual transformer models such as mT5, mBERT and other LLMs demonstrating competitive performance even in low resource languages like Bangla and Hindi [36].

Evaluation metrics for summarization are evolving from traditional n-gram overlap measures like ROUGE toward semantic and factual consistency assessments using neural natural language inference techniques and reference free metrics [37][38]. Such metrics exhibit higher correlation with human judgments, especially for abstractive and multilingual summaries.

Recent surveys of large language models (LLMs) emphasize their growing adaptability to diverse downstream tasks, including abstractive summarization in both high and low resource settings. Chang et al. (2024)[1] provide an extensive review of LLM evaluation strategies, highlighting challenges such as hallucination, factual drift and evaluation misalignment in summarization tasks. These concerns are especially relevant when deploying LLMs in sensitive domains like legal text generation. Similarly, Minaee et al. (2024)[2] trace the rapid evolution of generative transformer architectures and note the increasing reliance on fine-tuning and prompt engineering to adapt general purpose models for specialized domains. For Bangla legal texts, these trends underscore the need for more domain aware pretraining and evaluation frameworks that go beyond generic text generation metrics.

Complementary to model development, OCR quality significantly impacts downstream summarization outcomes, especially in languages with complex scripts like Bangla. Mukherjee and Saxena (2023)[6] discuss persistent challenges in OCR for handwritten Bangla documents, including irregular character spacing, complex ligatures and varied font styles. Studies like Saoji et al. (2021)[8] and Mukherjee et al. (2023)[6] demonstrate the effectiveness of pytesseract for Bangla script recognition, though they acknowledge accuracy drops in noisy or historical documents. Integrating such OCR pipelines with LLMs poses unique alignment challenges, as noisy input can degrade the semantic quality of generated summaries. These findings reinforce the importance of preprocessing, post correction and domain specific OCR tuning in developing reliable end-to-end summarization systems.

3.8 Research Gaps

Despite advancements, several key gaps remain in Bangla legal document summarization and OCR integration:

- The scarcity of large, annotated Bangla datasets for OCR and summarization limits supervised learning and model generalization.
- Existing models, including LLMs, lack domain specific fine-tuning on Bangla legal corpora, reducing summary accuracy and relevance.
- No comprehensive, publicly available pipeline integrates high accuracy Bangla OCR with abstractive summarization tailored specifically for land deed documents.

- Closing these gaps demands multidisciplinary collaboration in dataset creation, algorithm development and domain knowledge integration, leveraging modern transformer models and OCR advances to build practical, scalable solutions.
- Current research does not sufficiently address the accessibility and user interface challenges for non technical users (e.g., landowners, clerks, rural citizens) who would benefit the most from summarized deeds.
- General LLMs are not pretrained with legal specific objectives (e.g., recognizing clauses, obligations or property descriptors), leading to hallucinations or loss of legal nuance in Bangla summarization.
- Few works explore the joint optimization or robustness of summarization models in the presence of OCR induced noise common in scanned Bangla deeds (e.g., degraded print quality, old fonts, stamps)
- The field lacks publicly available benchmarks or evaluation datasets for Bangla deed summarization, limiting reproducibility and comparative analysis of different models.

Chapter 4

Dataset

4.1 Dataset Overview

The dataset used in this research consists of 477 Bangla land deed documents consisting of 4293 deed papers of data collected from government land offices and individual contributors. These documents include handwritten, printed and mixed formats, offering diversity in writing styles and layouts. The samples span across 9 different administrative regions in Bangladesh, covering deed records from around 2005 and represent various property types such as নাল, ভিটি, বাগান, জমি, চাষযোগ্য, বসতভিটা, ফাঁকা প্লট, কৃষি জমি. All documents were stored digitally in PDF, JPEG or PNG formats for consistency and ease of processing. This diverse dataset provided a robust foundation for OCR extraction and LLM-based summarization. The final version of our dataset was stored in a CSV file for further processing.

4.2 Data Collection

A foundational step in this research involved acquiring a dataset of Bangla land deeds suitable for both OCR processing and abstractive summarization. Given the sensitive nature of these documents, data collection was conducted through two primary channels:

i. Government Land Department Offices

We engaged with regional land offices to request access to deed documents. This route involved navigating administrative hurdles and maintaining strict ethical and privacy standards, as legal deeds often contain personal and financial information.

ii. Direct Collection from Individuals

To supplement the dataset and ensure diversity in handwriting, fonts and layout, personal land deeds from willing participants were collected. All contributors were informed about the purpose of the research and proper consent was taken to include their documents in the dataset. This helped us obtain a wider range of deed formats and qualities, increasing the robustness of our summarization model.

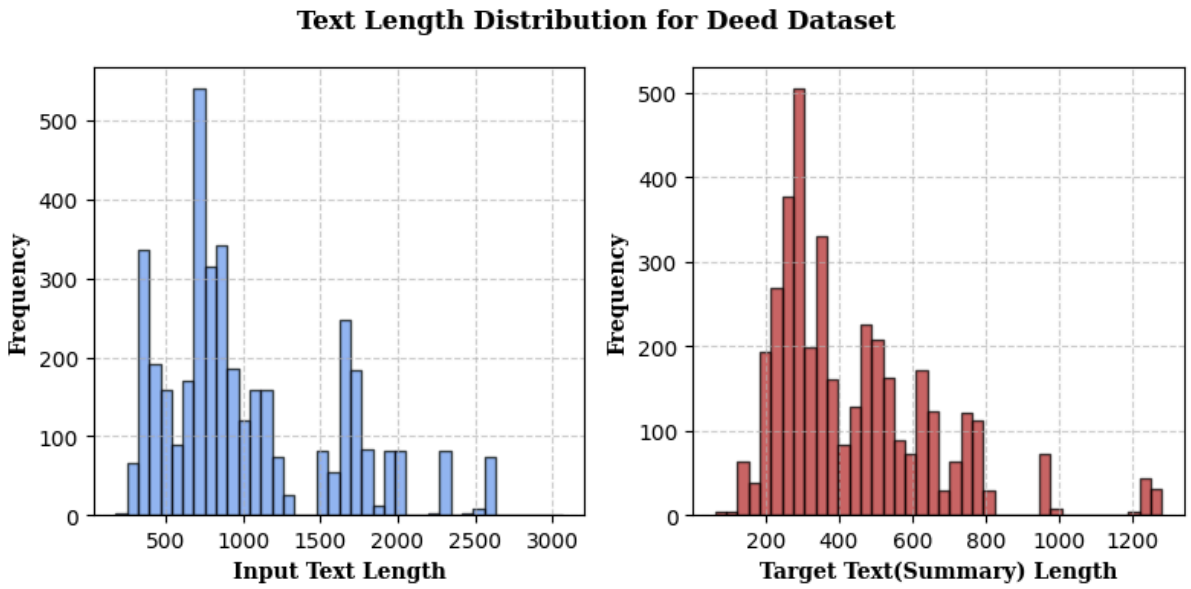


Figure 4.1: Text Length Distribution for Deed Dataset

4.3 Dataset Samples

4.3.1 Text extraction from a sample deed page

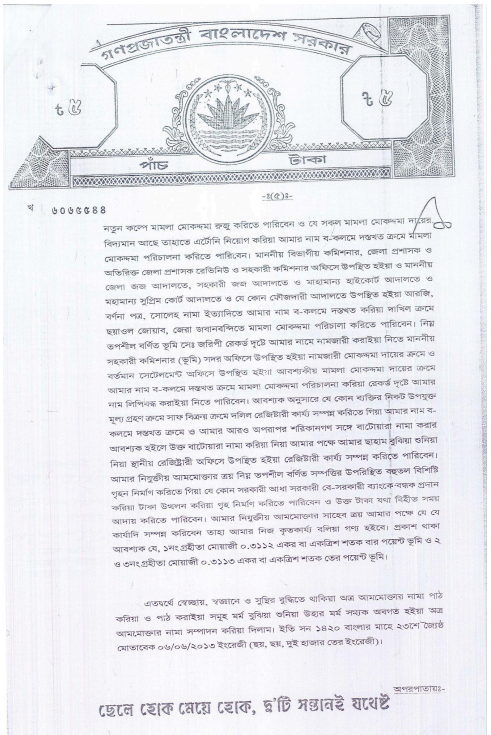


Figure 4.2: Sample Deed Page 1

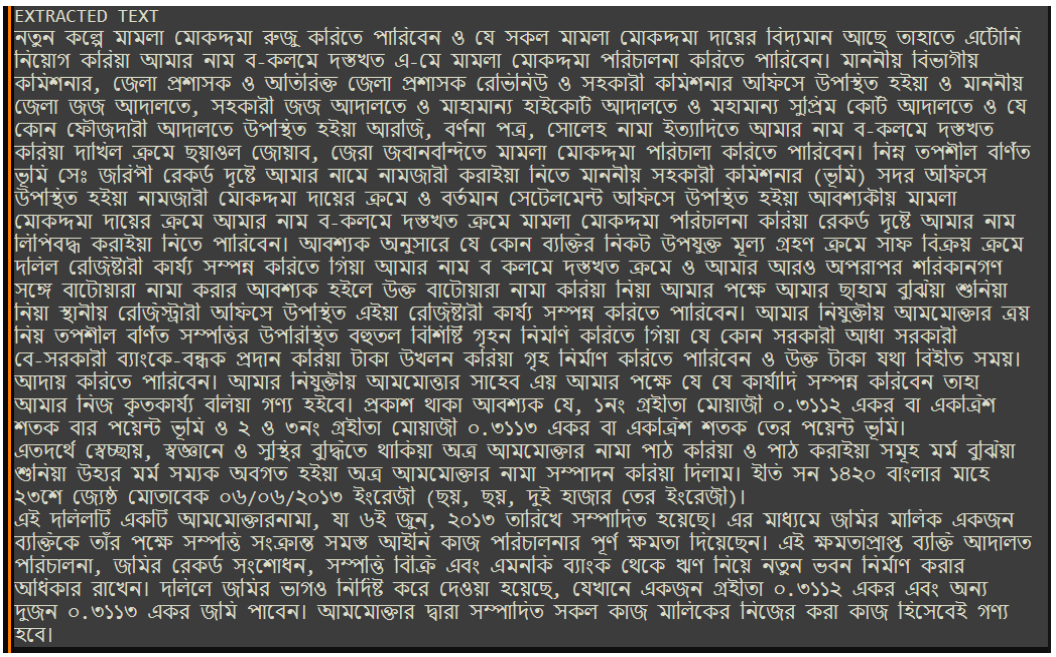


Figure 4.3: Extracted text or Input text

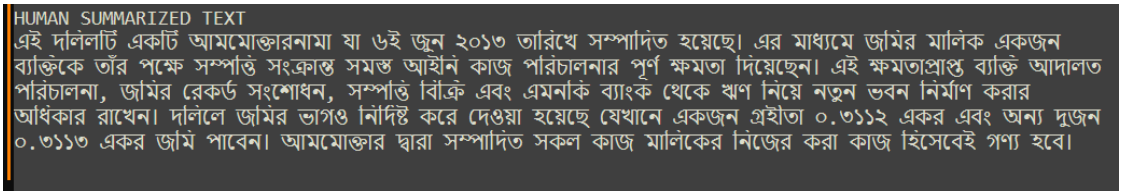


Figure 4.4: Manually Summarized Text

4.3.2 Some more samples from the deed collection

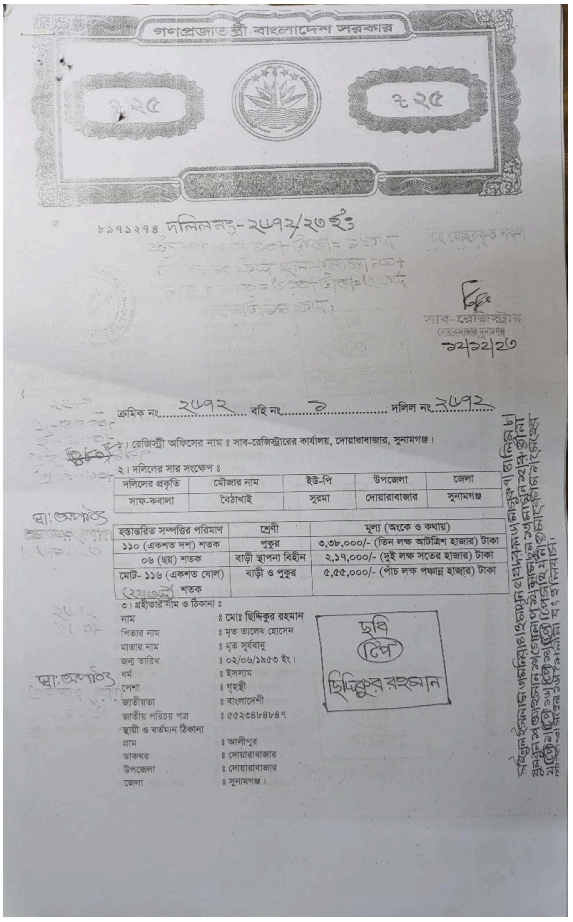


Figure 4.5: Sample Deed Page 2

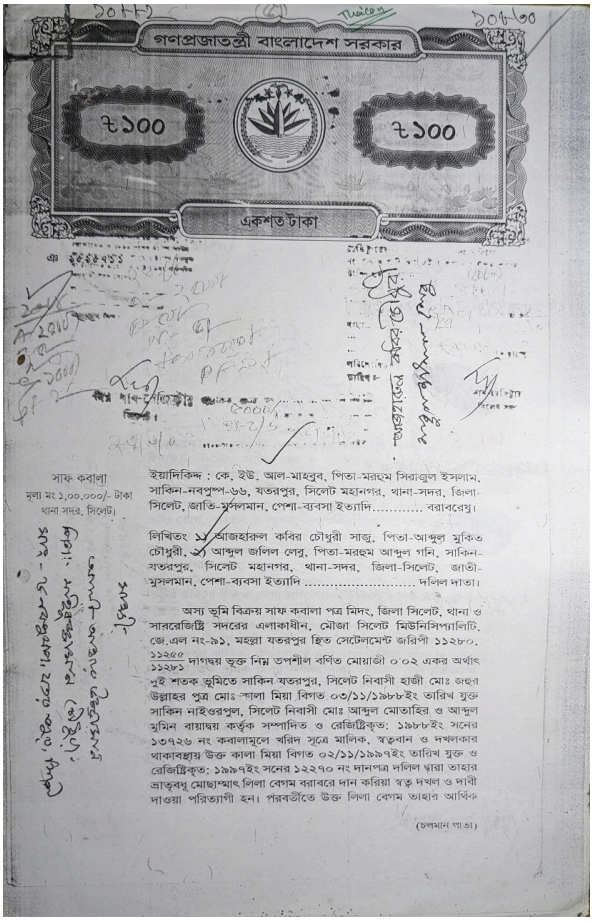


Figure 4.6: Sample Deed Page 3

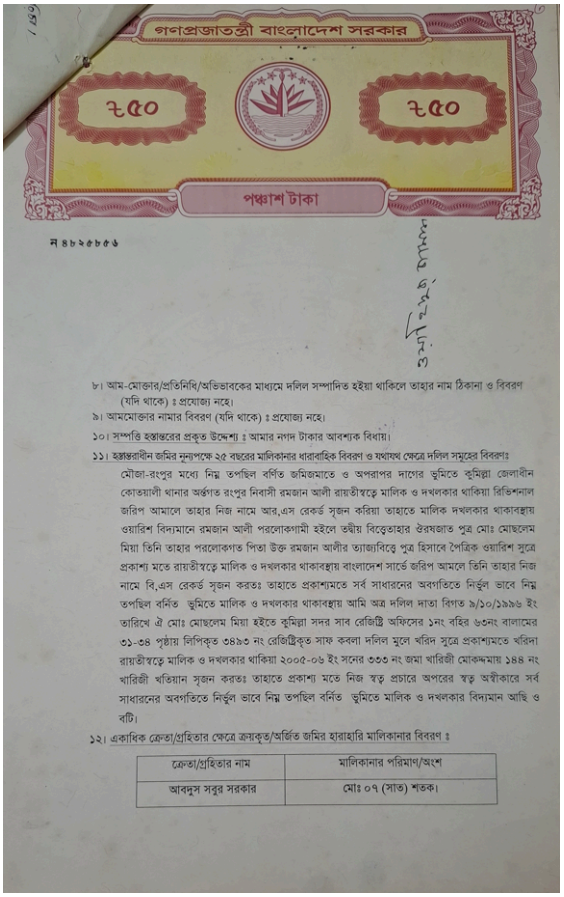


Figure 4.7: Sample Deed Page 4

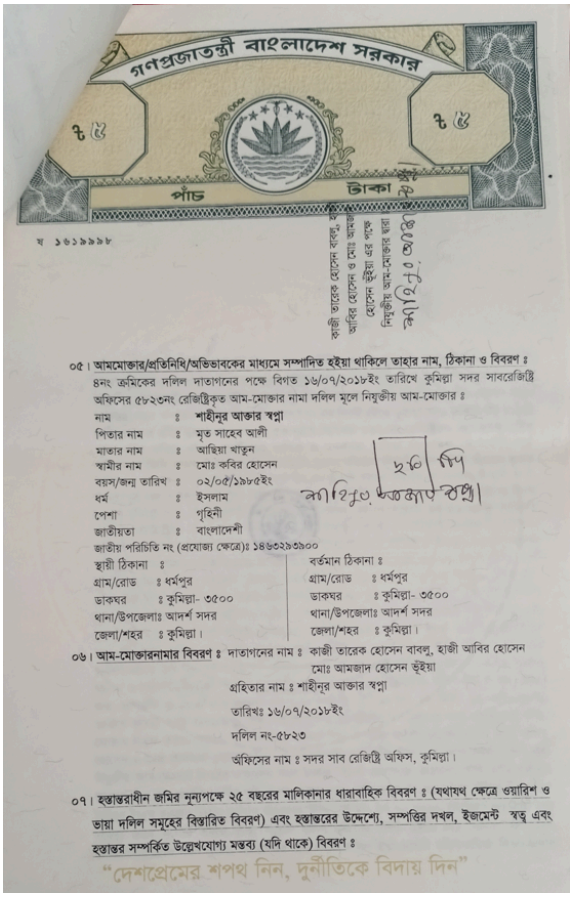


Figure 4.8: Sample Deed Page 5

4.3.3 Extracted text samples

কবির বিগত ২২/০১/১৯৮৭ইং তারিখে কুমিল্লা সদর জয়েন্ট সাবরেজিস্ট্রি অফিসের ৬০৫নং রেজিস্ট্রিকৃত সাফ কবলা দলিল মূলে সাকিন রাজাপুর নিবাসী রজ্জব আলীর পুত্র মোছলেম মিয়ার নিকট সাফ বিক্রয় করিয়া ভূমির দখল বুঝাইয়া দেয়। তৎপর মোছলেম মিয়া উক্ত বাবুল মিয়া, মাহাবুব আলম, আবদুল মতিন ও হুমায়ুন কবিরের নিকট হইতে খরিদ সূত্রে মালিক দখলকার থাকাবস্থায় বিগত ০৬/০৯/১৯৯৯ইং তারিখে কুমিল্লা সদর সাব রেজিস্ট্রি অফিসের ১নং বহির ৪৮নং বালামের ২৪৩ হইতে ২৪৬ পৃষ্ঠায় লিপিকৃত ৩২৭০নং রেজিষ্ট্রিকৃত সাফ কবলা দলিল মূলে আমি ১নং দলিল দাতার নিকট সাফ বিক্রয় করিয়া ভূমির দখল বুঝাইয়া দেয়। তৎপর আমি ১নং দলিল দাতা উক্ত মোছলেম মিয়ার নিকট হইতে খরিদ সূত্রে মালিক দখলকার হইয়া আমার নিজ নামে ১৯৯৯-২০০০ইং সনের ৫৯২নং জমা খারিজী মোকদ্দমায় ১৬৪নং খারিজা খতিয়ান সৃজনক্রমে ও বিগত ২৭/০৮/২০০৯ইং তারিখে উক্ত আইদর আলী ওরফে আয়দরের পুত্র ও কন্যা নাছির উদ্দিন, মোসাঃ নাজমা, এয়াছমিন এবং স্ত্রী সুফিয়া খাতুনের নিকট হইতে কুমিল্লা সদর সাব রেজিস্ট্রি অফিসের ৬৭৯৫নং রেজিস্ট্রিকৃত বি.এস খতিয়ানের ভুল সংশোধনের নাদাবী দলিল মূলে মালিক দখলকার হইয়া আমার নিজ নামে ২০০৬-০৭ইং সনের ১৩৪নং জমা খারিজী মোকদ্দমায় ৩১৬২২ খারিজা খতিয়ান সৃজনক্রমে এবং আমি ২নং দলিল দাত্রী বিগত ১৪/০৬/২০০৬ইং তারিখে কুমিল্লা সদর সাব রেজিষ্ট্রি অফিসের ৩৩৭১নং রেজিষ্ট্রিকৃত হেবার ঘোষণাপত্র দলিল মূলে আমার স্বামী অর্থাৎ ১নং দলিল দাতা আবদুল মান্নানের নিকট হইতে হেবার ঘোষণা প্রাপ্তে মালিক দখলকারিনী হইয়া আমার নিজ নামে ২০০৫-০৬ইং সনের ৩৯৮০নং জমা খারিজী মোকদ্দমায় ৩১২নং খারিজা খতিয়ান সৃজনক্রমে নিম্ন তপছিল বর্ণিত ভূমিতে মালিক দখলকার বিদ্যমান আছি। এইক্ষণ আমাদের নগদ টাকার বিশেষ প্রয়োজন হওয়ায় নিম্ন তপছিল বর্ণিত ভূমি সাফ বিক্রয় করার প্রস্তাব করিলে সেমতে আপনি দলিল গ্রহিত্রী উহা খরিদ করিতে ইচ্ছুক হওয়ায় তাহার বর্তমান বাজার দর উচিত মূল্য মং-১৪,০০,০০০/-(চৌদ্দ লক্ষ) টাকা মূল্য সাব্যস্থক্রমে মূল্যের সম্যক টাকা অদ্য আপনি গ্রহিত্রীর নিকট হইতে রোক নগদ গ্রহণ করিয়া আপনার নিকট নিম্ন তপছিল বর্ণিত ভূমি নির্দায় সাফ বিক্রয় করিলাম। বিক্রিত ভূমি অদ্যই আপনার দখলে বুঝাইয়া দিয়া আমরা আমাদের যৎযাবতীয় স্বত্ব লভ্য হইতে চিরতরে নিঃস্বত্ববান ও সর্বস্বত্ব ত্যাগী হইলাম। আপনি অদ্য হইতে বিক্রিত ভূমিতে আমাদের যাবতীয় স্বত্বে স্বত্ববান হইয়া দান, বিক্রয়, বিলএওয়াজ হেবা, হেবার ঘোষনাপত্র ইত্যাদি নানাহ প্রকার হস্তান্তরের অধিকারী হইয়া মালিক সরকারে আমাদের নামের পরিবর্তে আপনি আপনার নিজ নামে মালিক সরকারে নামজারী করতঃ সদর রাজস্ব আদায় পূর্বক আপনি ও আপনার পুত্র পৌত্রাদি অলি ওয়ারিশানগণক্রমে কাটিয়া ভরিয়া

Figure 4.9: Extracted text sample 1

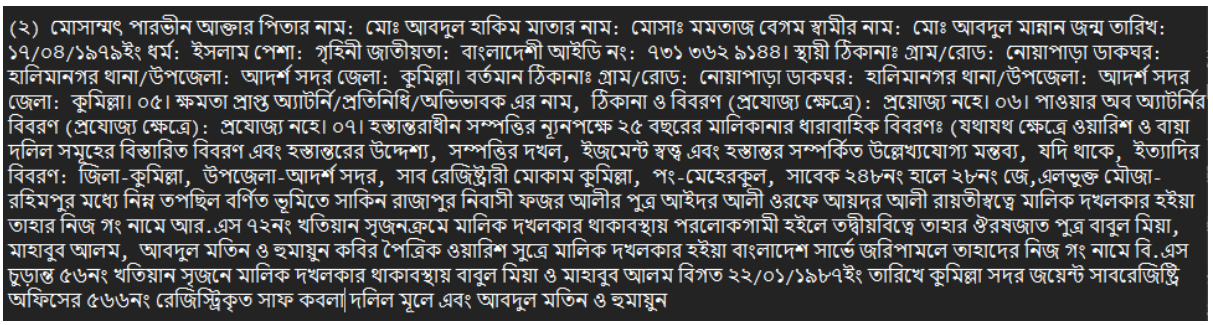


Figure 4.10: Extracted text sample 2

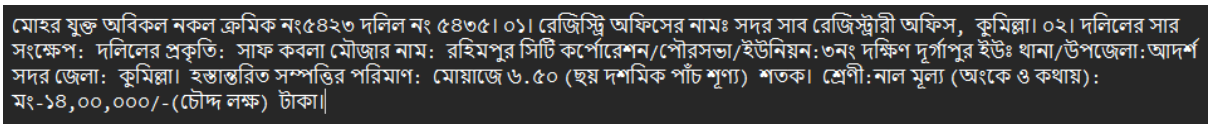


Figure 4.11: Extracted text sample 3

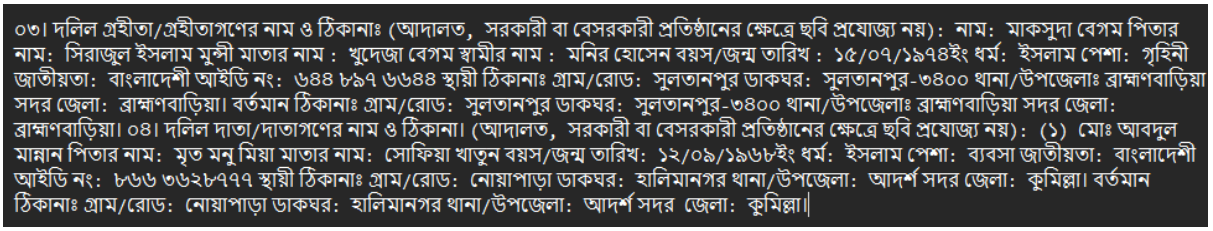


Figure 4.12: Extracted text sample 4

4.3.4 Summarized text samples

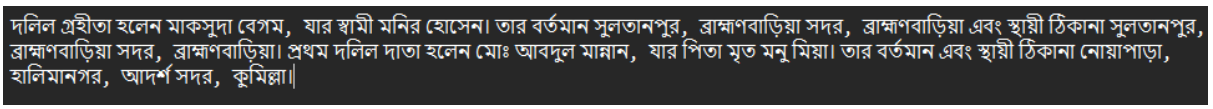


Figure 4.13: Summarized text sample 1

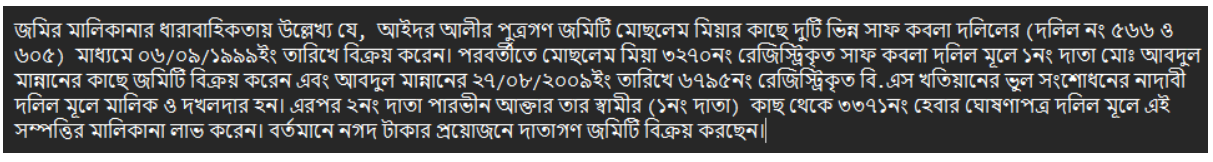


Figure 4.14: Summarized text sample 2

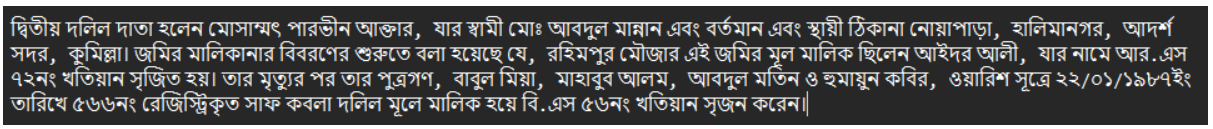


Figure 4.15: Summarized text sample 3

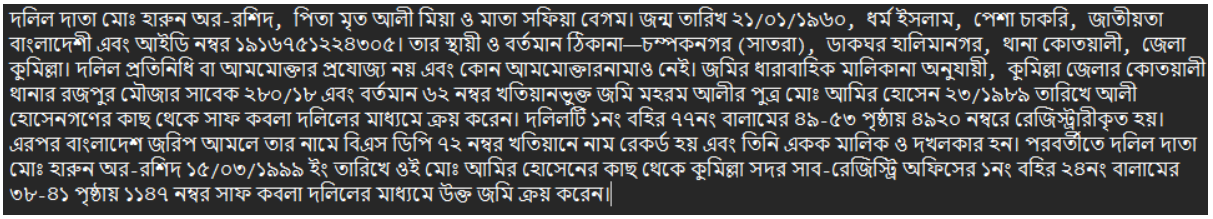


Figure 4.16: Summarized text sample 4

4.4 Preprocessing and Cleaning

The raw deed images collected were often noisy, misaligned or faint in text visibility. To ensure optimal OCR performance, a multi step preprocessing workflow was implemented. Standard denoising techniques were applied to eliminate background artifacts and remove scanning noise. Techniques such as contrast stretching, histogram equalization and sharpening filters were used to enhance faded text regions, enabling clearer segmentation by the OCR engine. Post OCR, the extracted text was often filled with recognition errors due to irregular fonts, handwritten annotations or scanning issues. A portion of these texts were manually corrected and used a rule based system to fix common OCR mistakes like similar looking Bangla characters misclassified by pytesseract.

This step significantly improved the quality of the extracted Bangla text, yielding cleaner, more accurate input for the downstream summarization model. It also ensured better sentence boundaries and token alignment during training.

4.5 Annotation Process

To create supervised training data for abstractive summarization, concise summaries of the extracted deed texts were manually written. Each summary was designed to capture the essential information such as parties involved, property details, deed number and registration location. This annotated dataset served as the ground truth for fine-tuning our language models.

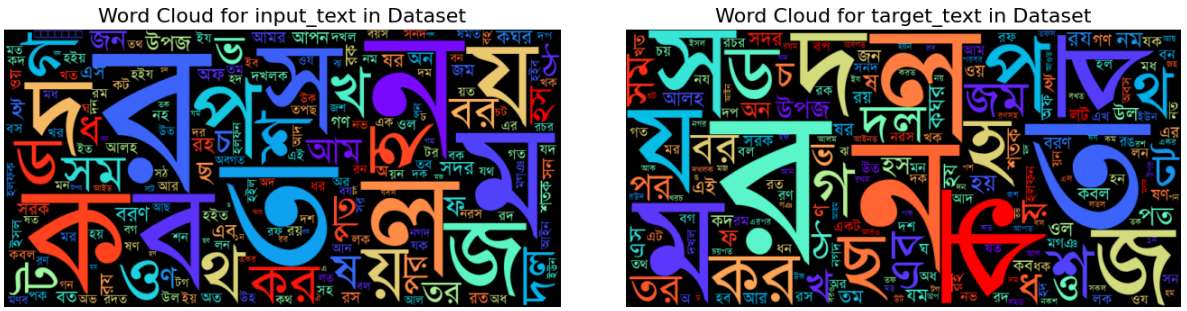


Figure 4.17: Word Cloud for input_text and target_text(summary) in Dataset

4.6 Ethical Considerations

Handling legal documents requires strict adherence to ethical standards to protect privacy and ensure informed consent. All personal land deeds used in this study were collected with explicit consent from the document owners, with full disclosure regarding the research purpose. Sensitive information such as names, addresses and deed numbers were anonymized during storage and processing. For government-sourced deeds, official permissions were obtained where required and data was handled in compliance with institutional and legal data protection standards. The dataset was not published publicly to preserve confidentiality.

Chapter 5

Methodology

This Chapter outlines the comprehensive methodology adopted to work towards the abstractive summarization of Bangla land deeds. The workflow integrates OCR based text extraction with fine-tuned language models for generative summarization, following a structured approach comprising data collection, preprocessing, extraction and labelling, model selection and adaptation, summarization and evaluation. Each step has been tailored to address the unique challenges posed by Bangla legal documents.

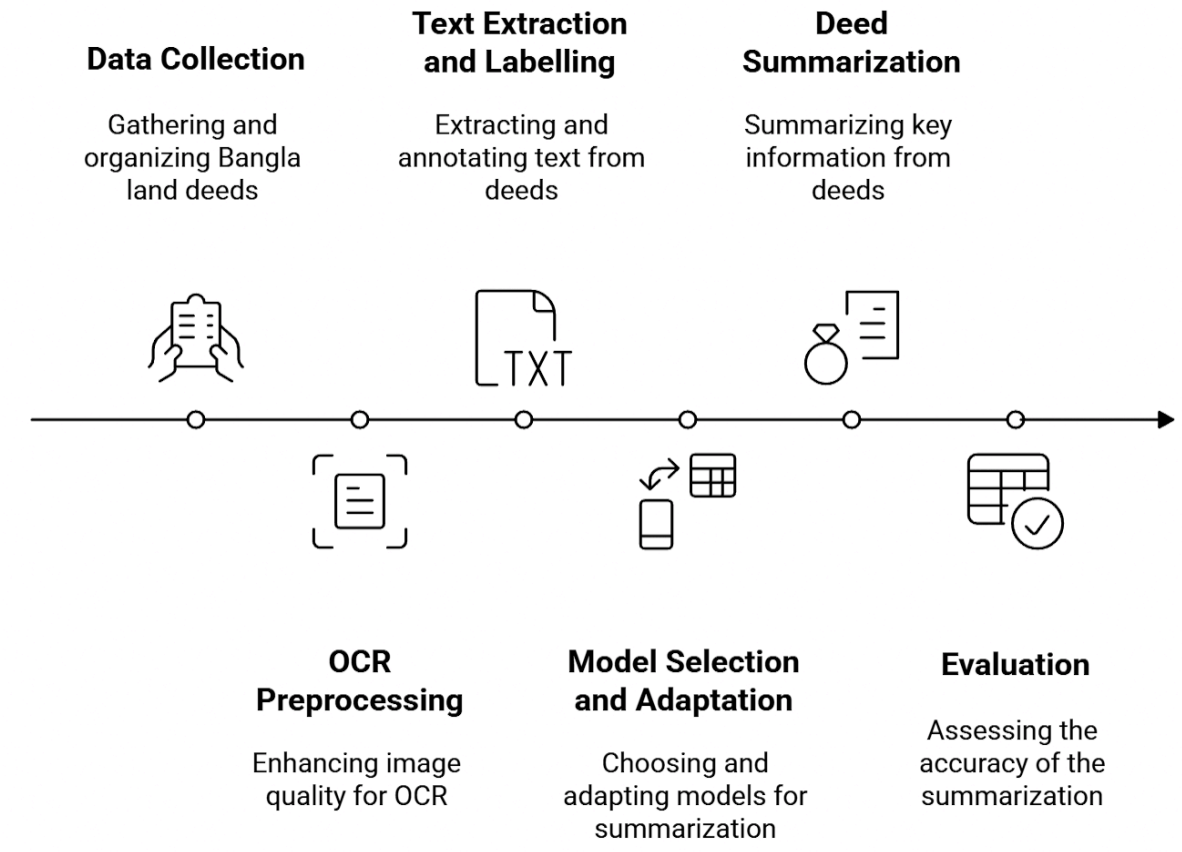


Figure 5.1: Methodology Overview

5.1 Data Overview

To support the summarization task, a curated dataset of Bangla land deeds was utilized, as detailed in Chapter 4. These documents were obtained from both government offices and private contributors. Ethical considerations and consent protocols were strictly followed. The dataset forms the backbone of our OCR and summarization pipeline.

5.2 OCR Preprocessing

Preprocessing steps including denoising, binarization and skew correction were applied to enhance image quality and OCR performance. These techniques were used to improve text clarity for downstream tasks. Full preprocessing procedures are detailed in Chapter 4.

5.3 Text Extraction and Summary Modeling

Text was extracted from deed images using Tesseract-based OCR tuned for Bangla script. Following extraction, fine-tuned LLMs were employed to generate abstractive summaries. Given summaries were manually annotated as described in Chapter 4, enabling supervised training of our models.

5.4 Model Selection and Adaptation

This step focused on identifying suitable large language models (LLMs) and adapting them to the specific requirements of Bangla legal text summarization.

i. Model Selection

Given the limited availability of pre-trained Bangla specific LLMs, we worked with multilingual transformer based models like mT5 and banglaT5, known for their capacity to handle low resource languages. These models were selected due to their generative capabilities and strong performance in sequence-to-sequence tasks.

ii. Adaptation

The chosen models are fine-tuned using our custom labelled dataset. This involved formatting the inputs and outputs in a standardized structure, adjusting the maximum sequence lengths and modifying tokenizers to better handle Bangla script. Experimentation with multiple hyperparameter settings such as learning rate, batch size, gradient evaluation steps and learning rate scheduler type to optimize the summarization output.

iii. Outcome

The fine-tuned models demonstrated the ability to generate coherent, domain relevant summaries from our given test data. Adaptation also reduced hallucination and improved fidelity to the source content.

5.5 Deed Summarization

The core of our methodology involved implementing and refining the abstractive summarization module using the adapted LLMs.

i. Implementation

We designed a pipeline where OCR extracted text was passed into the fine-tuned LLM, which then generated a compact summary. We highlighted some key components from our test data to be available in the summary like names, locations, plot descriptions and deed types, which helped us build more concise summaries.

ii. Optimization

We optimized the summarization model through extensive hyperparameter tuning, including learning rate, batch size and token lengths. A custom AdamW optimizer and learning rate scheduler as cosine with restarts were employed to ensure stable convergence. We used early stopping to prevent overfitting and checkpointing to resume interrupted training. The evaluation and saving were aligned per epoch and a custom data collator was implemented to handle padded tensor batches efficiently.

5.6 Evaluation

A robust evaluation framework was essential to assess the performance, relevance and reliability of our Bangla land deed summarization system.

i. Evaluation Metrics

We utilized both lexical and semantic metrics to evaluate model-generated summaries. Key metrics included:

- Character Error Rate (CER) and Word Error Rate (WER) to measure textual deviation from the reference summaries.
- Exact Match (EM) to assess the proportion of predictions that exactly matched given summaries.
- BERTScore, using bert-base-multilingual-cased with Bengali language support, to evaluate deep semantic similarity between generated and reference texts. This metric helped assess the model's ability to retain the meaning and context of legal clauses despite lexical variations.

ii. Evaluation Setup

Summaries were generated from the test set using beam search with a length penalty and all tensors were processed on GPU for faster inference. Predictions and references were decoded using the tokenizer and metrics were computed in batches. Text normalization (e.g., Unicode simplification) was applied where necessary for fair comparison.

iii. Outcome

After multiple training sessions, the models mT5 and banglat5 achieved a high BERTScore F1 ranging between 0.96 to 0.97, indicating strong semantic alignment with ground truth. CER and WER remained low (approximately 9 to 12% and 12 to 15% respectively), reflecting good surface-level accuracy. Exact Match was 50% to 57%, demonstrating the models capability to generate precise summaries in many cases. These results confirm that fine-tuning on domain specific legal data significantly enhances both factual correctness and semantic coherence. The evaluation also revealed some performance limitations with complex or underrepresented deed structures, guiding future areas for model improvement.

5.7 Workstation Configuration

To train and evaluate the mT5 and BanglaT5 models used in this research two high-performance desktop workstations were utilized. Both systems were equipped with NVIDIA RTX 3060 GPUs (12GB VRAM), enabling the use of CUDA acceleration through the PyTorch framework for efficient deep learning training.

The environment on both machines included Python 3.11.9 and PyTorch 2.6.0, which ensured compatibility with the latest features and CUDA optimizations available for GPU-based model training.

System Specifications:

System 1 (used for training mT5)

- Processor: AMD Ryzen 7 5800X — 8 cores / 16 threads
- GPU: NVIDIA RTX 3060 12GB
- RAM: 32GB DDR4
- Operating System: Windows 11

System 2 (used for training BanglaT5)

- Processor: Intel Core i5-12600K — 10 cores / 16 threads
- GPU: NVIDIA RTX 3060 12GB
- RAM: 16GB DDR4
- Operating System: Windows 11

Performance Observation:

During experimentation, System 1 consistently delivered better training throughput and stability compared to System 2. Despite the Intel i5-12600K's more physical cores and multi-threaded performance, the system with Ryzen 7 5800X and larger RAM capacity (32 vs 16) significantly improved training time. This minimized memory related bottlenecks, especially for larger models like mT5, which require substantial resources for effective fine-tuning. As a result, mT5 was trained on System 1, whereas the relatively lighter BanglaT5 was trained on System 2.

This configuration balance ensured optimal utilization of available hardware resources without compromising model performance or stability during training.

5.8 Experimental Setup and tools used

The experimental setup of this research integrated several components across model fine-tuning, evaluation and visualization workflows. The process was executed using Python, with a combination of open-source libraries and frameworks tailored for legal text summarization in Bangla.

Modeling Frameworks

- Transformers (via Hugging Face): Used to fine-tune mT5 and BanglaT5 models on the custom dataset.
- Datasets: For loading and structuring the training, validation and test splits.
- PyTorch: Backend deep learning framework supporting training routines and GPU acceleration.
- AdamW Optimizer, learning rate scheduler type as cosine with restarts and early stopping were employed to ensure stable and efficient convergence.

Evaluation Tools

- BERTScore: To measure semantic similarity using bert-base-multilingual-cased.
- JIWER: To compute Word Error Rate (WER).
- Evaluate: Hugging Face's library for standard metrics like BLEU and Exact Match.
- Unicode: Used for normalization and preprocessing of Bangla text prior to metric calculation.

Visualization and Analysis

- Matplotlib: Used to plot training and validation loss curves.
- Wordcloud: Employed to visually analyze common terms in generated summaries.

- Pandas: For data manipulation, tabulation of metrics and logging experimental results.
- Scikit-learn: Used for basic statistical analysis and support functions during evaluation.

Execution Environment

Training and evaluation were performed using NVIDIA GPUs to accelerate computation. Beam search decoding (num_beams=4, length_penalty=2.0) was applied during inference to enhance summary quality. All results were recorded for reproducibility and graphs and logs were preserved.

Chapter 6

Results and Discussion

This Chapter presents a comprehensive evaluation of the proposed abstractive summarization models mT5 and BanglaT5. These models are fine-tuned on a custom dataset of Bangla land deeds. We analyze model performance across multiple dimensions using a combination of surface level, semantic and structural metrics, followed by a comparative discussion on the suitability of each model for Bangla legal text summarization.

6.1 Evaluation Metrics and Strategy

To comprehensively assess the quality of the generated summaries, we adopted both automatic evaluation metrics and human interpretable insights. These metrics were chosen to reflect the diverse requirements of legal summarization, including factual correctness, fluency, semantic alignment and structural faithfulness.

Metrics used:

Character Error Rate (CER): Measures the proportion of character-level edits needed to convert the predicted summary into the reference.

Word Error Rate (WER): Similar to CER but at the word level. This is especially relevant for legal texts where word choice is critical.

Exact Match (EM): Indicates the percentage of summaries that exactly match the reference, capturing precision in legal terminology.

BLEU Score: Measures n-gram overlap between predicted and reference summaries; commonly used in machine translation and summarization tasks.

BERTScore: Uses contextual embeddings to assess semantic similarity, especially important for abstractive tasks involving paraphrasing.

These metrics collectively provide a robust perspective on both the factual fidelity and linguistic coherence of the summaries. Evaluation was performed on the held-out test set using beam search (num_beams=4, length_penalty=2.0) for decoding.

6.2 Quantitative Results

Metric	mT5	BanglaT5
Character Error Rate (CER)	9.97%	12.08%
Word Error Rate (WER)	12.34%	15.02%
Exact Match (EM)	50.91%	56.14%
BLEU Score	0.905	0.886
BERTScore (Precision)	0.9666	0.9641
BERTScore (Recall)	0.9588	0.9575
BERTScore (F1)	0.9624	0.9605

Table 6.1: Quantitative Results of mT5 and BanglaT5

Figure 6.1: Model Performance Comparison Across Metrics

6.3 Analysis and Interpretation

6.3.1 Surface-Level Accuracy: CER and WER

The mT5 model exhibited lower CER (9.97%) and WER (12.34%) values, reflecting a higher degree of textual precision. Generated summaries typically maintained fluent, formal Bangla but occasionally generalized specific details like names or monetary figures. BanglaT5, on the other hand, tended to paraphrase more freely—causing slightly higher surface-level error rates.

6.3.2 Token-Level Precision: Exact Match

BanglaT5 achieved higher exact match accuracy, indicating stronger alignment with the original phrasing and terminology. Many BanglaT5 summaries preserved critical legal vocabulary (e.g., "হস্তান্তরিত", "উত্তরাধিকার", "দলিল গ্রহীতা") with higher fidelity than mT5, which sometimes introduced slightly altered phrases.

6.3.3 Semantic Quality: BERTScore

Both models performed strongly on semantic alignment. Even when wording varied, the generated content from each model preserved core meanings. The mT5 summaries favored smoother sentence flow, while BanglaT5 more consistently emphasized structured legal context.

6.3.4 N-Gram Overlap: BLEU

The BLEU score was slightly higher for mT5 (0.905), indicating better lexical similarity with reference summaries. This suggests that mT5 favored a more conservative summarization style, reusing phrases from the source with higher fidelity. Qualitative inspection supports

this, as mT5 produced more readable, concise summaries. In contrast, BanglaT5 (0.886) generated denser, detail-rich outputs, often emphasizing critical elements such as party names, property details and registration information—even when expressed with varied phrasing.

6.4 Model Behavior Observations

Based on sample-wise qualitative assessment:

- mT5 generated grammatically fluent summaries with a neutral tone. However, it sometimes omitted or abstracted key legal identifiers like land type or land value.
- BanglaT5 often retained formal legal expressions and complete names of parties, making its summaries more suitable for legal review. However, occasional repetitions or OCR-induced distortions were observed.
- Both models struggled when source texts had OCR artifacts, underlining the importance of robust input cleaning.

6.5 Visual Analysis

This Chapter presents an in-depth analysis of the training dynamics of both the mT5 and BanglaT5 models, based on complete epoch-wise training and validation losses over 30 epochs. These metrics not only reflect the learning behavior of each model but also provide insights into optimization stability, convergence efficiency and potential overfitting. Graphical visualizations (loss_curve.pdf) and detailed logs (loss_history.csv) have been preserved in the appendix for transparency and reproducibility.

6.5.1 BanglaT5 Loss Trends

The BanglaT5 model exhibited a strong and consistent decline in both training and validation losses over 30 epochs:

- Training loss decreased from 3.46 (Epoch 1) to 0.0287 (Epoch 30), an impressive ~99% reduction.
- Validation loss followed a smooth trajectory, dropping from 2.93 to 0.0236, showing strong generalization performance.

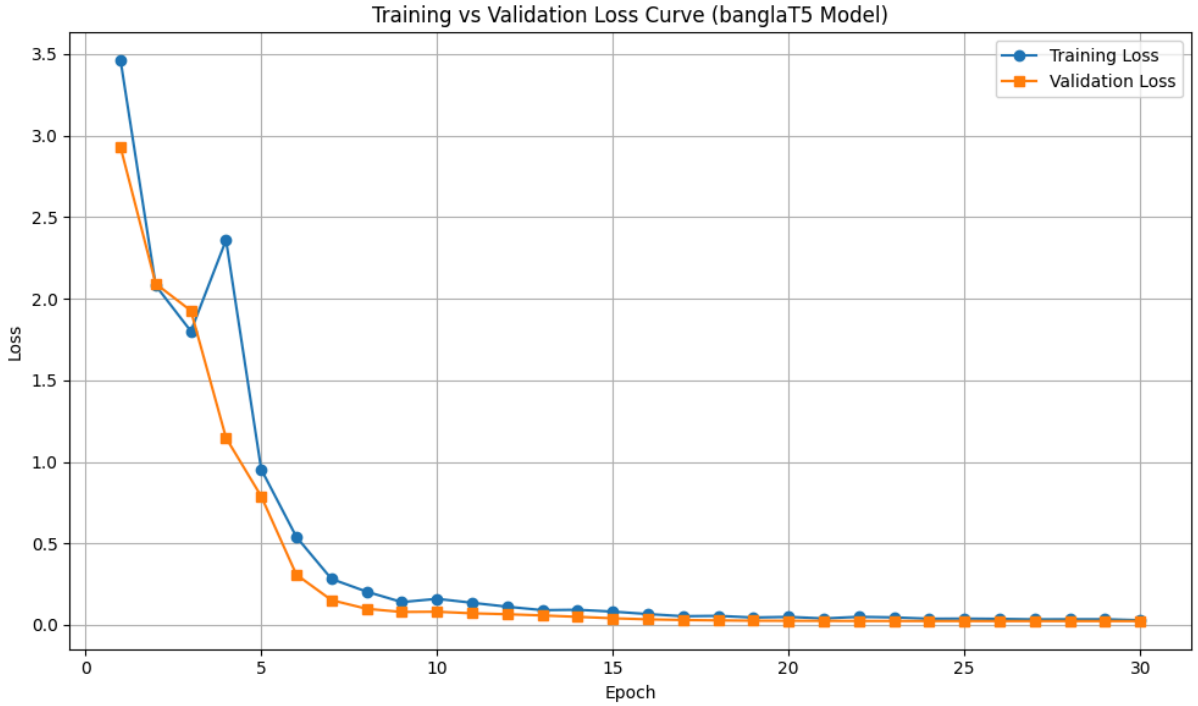


Figure 6.2: Training vs Validation Loss Curve (banglaT5 Model)

Notably, the most substantial performance gains occurred within the first 10 epochs, during which validation loss plummeted from 2.93 to 0.0808. After this point, the loss curve continued to decline steadily, plateauing after Epoch 20 at approximately 0.0236.

This behavior highlights two key observations:

- Fast convergence: BanglaT5 adapted to the legal domain rapidly, benefiting from its monolingual pretraining on Bangla text.

- Stable optimization: No signs of overfitting were observed even in later epochs, thanks to proper regularization, early stopping safeguards and learning rate scheduling.

The model maintained tight coupling between training and validation losses throughout, suggesting a high quality fit on the data with strong generalization to unseen examples.

6.5.2 mT5 Loss Trends

The mT5 model also demonstrated strong convergence across epochs but with a more gradual progression compared to BanglaT5:

- Training loss dropped from 4.58 (Epoch 1) to 0.0371 (Epoch 30).
- Validation loss improved from 4.23 to 0.0329, demonstrating excellent generalization and learning stability.

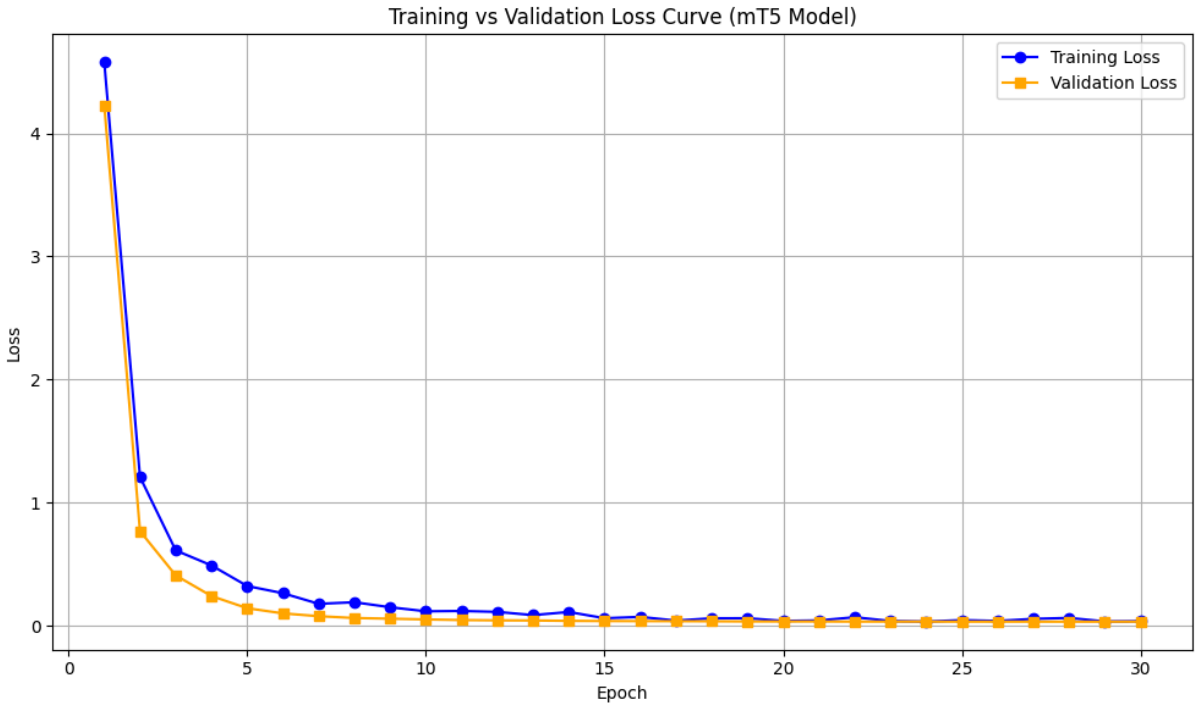


Figure 6.3: Training vs Validation Loss Curve (mT5 Model)

Key characteristics of mT5's training curve:

- Initial training was slightly noisier, with a steeper early drop (Epochs 1–5), followed by a slower but consistent refinement phase from Epochs 10 to 30.
- From Epoch 8 onward, validation loss dropped below 0.07 and gradually converged toward ~0.0328, indicating a reliable and stable learning process.
- Slight fluctuations (e.g., at Epochs 14, 18, 27) are expected in such low resource domains and are mitigated by appropriate learning rate restarts.

Despite being a multilingual model not originally optimized for Bangla, mT5 managed to reach performance comparable to BanglaT5, a testament to the model's cross lingual capacity and the strength of the fine-tuning setup.

6.5.3 Comparative Visualization and Interpretation

The following table summarizes key training dynamics:

Metric	mT5	BanglaT5
Initial Validation Loss	4.228	2.928
Final Validation Loss	0.0329	0.0236
Total Drop in Val. Loss	-4.195	-2.904
Overfitting Observed	No	No
Convergence Speed	Gradual	Faster
Training Stability	High	Very High

Table 6.2: Comparative Visualization and Interpretation of mT5 and BanglaT5

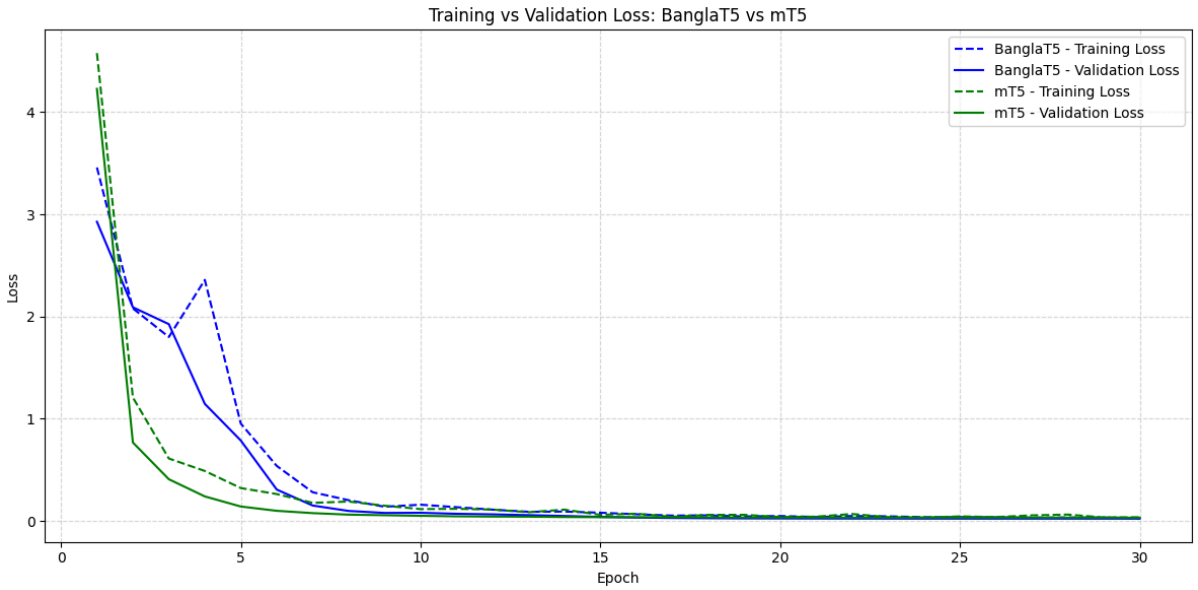


Figure 6.4: Training vs Validation Loss: BanglaT5 vs mT5

Both models demonstrated healthy convergence and effective generalization. However, BanglaT5 outperformed mT5 in convergence speed and final validation loss, attributable to its Bangla centric pretraining. In contrast, mT5 excelled in stability, showcasing its adaptability despite being trained across multiple languages.

6.5.4 Key Insights

The cosine learning rate scheduler with restarts, coupled with **e**arly stopping and AdamW optimization, provided robust training stability for both models.

BanglaT5 is more sample efficient, learning complex legal semantics in fewer epochs with faster loss reduction.

mT5 remains a strong baseline, offering comparable final performance and useful for multilingual or zero-shot extensions.

No divergence or overfitting was detected in either model, validating the effectiveness of the chosen hyperparameters and training strategy.

In conclusion, the loss curve analysis reaffirms that both models are well optimized for Bangla deed summarization, with BanglaT5 holding a marginal edge in convergence and loss

minimization. These observations provide a strong foundation for selecting and deploying models in real world applications where accuracy, fluency and reliability of summaries are critical.

6.6 Summary of Comparative Performance

Aspect	mT5	BanglaT5
Language Adaptability	High (Multilingual)	High (Bangla-focused)
Surface Accuracy	Lower error rates	Moderate error rates
Exact Matching	Moderate	Higher match rate
Semantic Similarity	High (BERTScore: 0.962)	High (BERTScore: 0.960)
Domain Fluency	Good	Excellent in legal Bangla context

Table 6.3: Comparative Performance of mT5 and BanglaT5

6.7 Implications for Legal Text in Bangla

Both models are viable for Bangla land deed summarization: mT5 is suitable for applications needing fluent, human readable summaries for public use. BanglaT5 better supports legal professionals requiring accurate and structured paraphrases of original deeds. These findings support the hypothesis that domain specific fine-tuning, especially in low resource languages, plays a critical role in real world legal AI systems.

6.8 Model Determining Table

Aspect	Best Performing Model	Interpretation
Semantic Similarity (F1)	mT5	Higher fluency and paraphrasing quality
Exact Phrase Accuracy (EM)	BanglaT5	Better preservation of legal phrasing
Content Coverage	BanglaT5	Better structural alignment with key information
Surface level Accuracy (CER/ WER)	mT5	Fewer character and word level deviations
Overall Legal Suitability	Balanced	mT5 for fluency; BanglaT5 for legal precision

Table 6.4: Model Determining Table

Chapter 7

Conclusion and Future Work

7.1 Conclusion

This research presented a comprehensive pipeline for the abstractive summarization of Bangla land deeds, integrating Optical Character Recognition (OCR) with fine-tuned transformer-based language models. By creating a high quality dataset of legal documents and their summaries, we addressed a significant gap in legal informatics for low-resource languages like Bangla.

Two models, the multilingual mT5 and the monolingual BanglaT5, were trained and evaluated. The results revealed a crucial trade-off: mT5 produced more fluent and generalized summaries with lower surface-level errors, whereas BanglaT5 excelled at preserving critical domain-specific terminology, legal phrasing and structural components such as names and property details. The superior performance of BanglaT5 in capturing legal nuances highlights the benefits of monolingual, domain-specific pretraining for specialized tasks.

Ultimately, this study demonstrates the feasibility and value of applying fine-tuned language models to enhance the accessibility of complex legal documents. The findings lay a strong foundation for practical applications in legal aid, registry automation and accessible legal technology for Bangla speakers, contributing valuable insights to the intersection of Bangla NLP and legal document understanding.

7.2 Future Work

This study provides a strong proof-of-concept for summarizing complex Bangla legal documents using modern NLP techniques. However, this is an initial step into a broad and impactful research area. To build upon this foundation, several future directions can be pursued to enhance the system's accuracy, robustness and real world utility. The subsequent sections detail potential improvements related to data curation, model architecture, human centric evaluation and practical deployment.

7.2.1 Dataset Expansion and Diversification

Future research should focus on creating a larger and more diverse corpus by collecting various deed types and regional formats. A more comprehensive dataset will be crucial for improving model generalization and would enable the system to better accommodate historical documents and evolving legal terminology.

7.2.2 Enhancing Robustness to OCR Errors

An important area for future research will be to develop OCR aware summarization models that are robust to noisy input. This would involve training the model to tolerate or correct common OCR errors, which should ensure more reliable performance with imperfectly scanned, real-world documents.

7.2.3 Legal Domain Continued Pretraining

A key direction for future work will be to perform continued pretraining of the base language model on a large corpus of unlabeled Bangla legal texts. This domain adaptation would enrich the model's understanding of legal vocabulary and syntax, which should lead to more factually grounded and legally faithful summaries.

7.2.4 Rigorous Human-in-the-Loop Evaluation

Beyond automated metrics, it will be essential to conduct a structured evaluation involving legal professionals to assess the practical utility of the summaries. Such a process should use a framework to evaluate legal correctness and factual accuracy, which would provide invaluable insights for real world deployment.

7.2.5 Minimizing Legal Hallucinations

Given the critical nature of legal documents, even minor hallucinations or omissions in summaries can result in misleading interpretations. While current evaluation metrics such as BERTScore indicate high semantic alignment, they do not always capture factual fidelity, especially in domain specific contexts like land ownership, plot boundaries or legal clauses. To mitigate this, future work should focus on integrating factual consistency checks, clause

specific labeling tasks, using contrastive learning or instruction tuning with legal constraints. Developing legal domain reward models to discourage hallucinations during generation can also assist in this regard.

7.2.6 Practical Deployment and User Access

To make this research usable by the public, a deployment plan can be devised to convert the summarization pipeline into an accessible tool. This could take the form of a web or mobile application, enabling users to upload scanned deeds and receive summaries in plain Bangla. The tool should feature an OCR preprocessing module with visual feedback, a confidence score for summaries, optional human review for critical decisions and voice assistance for accessibility in rural areas. Besides, it can also target users including landowners, registry clerks and legal aid centers, especially in underserved rural regions. Future work should also explore collaborations with local government portals or NGOs to ensure equitable access.

References

[1] Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., ... & Xie, X. (2024). A survey on evaluation of large language models. *ACM transactions on intelligent systems and technology*, *15*(3), 1-45.

[2] Minaee, S., Mikolov, T., Nikzad, N., Chenaghlu, M., Socher, R., Amatriain, X., & Gao, J. (2024). Large language models: A survey. *arXiv preprint arXiv:2402.06196*.

[3] Sen, O., Fuad, M., Islam, M. N., Rabbi, J., Masud, M., Hasan, M. K., ... & Iftee, M. A. R. (2022). Bangla natural language processing: A comprehensive analysis of classical, machine learning, and deep learning-based methods. *IEEE Access*, *10*, 38999-39044.

[4] Rao, D. R., Harika, B., Srikanth, A., & Vahini, Y. (2022, February). Nlp based machine learning approaches for text summarization. In *Proceedings of the International Conference on Innovative Computing & Communication (ICICC)*.

[5] Divya, K., Sneha, K., Sowmya, B., & Rao, G. S. (2020). Text summarization using deep learning. *International Research Journal of Engineering and Technology (IRJET)*, *7*(05), 3673-3677.

[6] Mukherjee, P., & Saxena, A. (2023). Challenges and solutions in OCR of handwritten Bangla documents. *Journal of Namibian Studies*, 35, 283–292.

[7] Isheawy, N. A. M., & Hasan, H. (2015). Optical character recognition (OCR) system. *IOSR Journal of Computer Engineering (IOSR-JCE), e-ISSN*, 2278-0661.

[8] Saoji, S., Eqbal, A., & Vidyapeeth, B. (2021). Text recognition and detection from images using pytesseract. J Interdiscip Cycle Res, 13, 1674-1679.

[9] John, I. (2025). Fine-Tuning LLMs for User Intent: A Framework for Contextualized Text Summarization.

[10] Rehman, T., Ghosh, S., Das, K., Bhattacharjee, S., Sanyal, D. K., & Chattopadhyay, S. (2025). Evaluating LLMs and Pre-trained Models for Text Summarization Across Diverse Datasets. *arXiv preprint arXiv:2502.19339*.

[11] Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E. D., Gutierrez, J. B., & Kochut, K. (2017). Text summarization techniques: a brief survey. *arXiv preprint arXiv:1707.02268*.

[12] Abujar, S., Hasan, M., & Hossain, S. A. (2018, October). Sentence similarity estimation for text summarization using deep learning. In *Proceedings of the 2nd International Conference on Data Engineering and Communication Technology: ICDECT 2017* (pp. 155-164). Singapore: Springer Singapore.

[13] Shakil, H., Farooq, A., & Kalita, J. (2024). Abstractive text summarization: State of the art, challenges, and improvements. *Neurocomputing*, *603*, 128255.

[14] Zhang, M., Zhou, G., Yu, W., Huang, N., & Liu, W. (2022). A comprehensive survey of abstractive text summarization based on deep learning. *Computational intelligence and neuroscience*, *2022*(1), 7132226.

[15] Arifuzzaman, M., & Islam, S. (2024). Digitalization of Land Documents in Bangladesh: Challenges and Prospects. *Open Access Library Journal*, *11*(10), 1-9.

[16] Farzindar, A., & Lapalme, G. (2004, July). Legal text summarization by exploration of the thematic structure and argumentative roles. In *Text Summarization Branches Out* (pp. 27-34).

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

[18] Sultana, F., Fuad, M. T. H., Fahim, M., Rahman, R. R., Hossain, M., Amin, M. A., ... & Ali, A. A. (2024, December). How Good are LM and LLMs in Bangla Newspaper Article Summarization?. In *International Conference on Pattern Recognition* (pp. 72-86). Cham: Springer Nature Switzerland.

[19] Borah, M., Dadure, P., & Pakray, P. (2022). Comparative analysis of T5 model for abstractive text summarization on different datasets.

[20] Miazee, A. A., Roy, T., Islam, M. R., & Safat, Y. (2025, February). Abstractive Text Summarization for Bangla Language Using NLP and Machine Learning Approaches. In *2025 International Conference on Electrical, Computer and Communication Engineering (ECCE)* (pp. 1-4). IEEE.

[21] Rahman, A., Rafiq, F. M., Saha, R., & Rafian, R. (2018). *Bengali text summarization using TextRank, Fuzzy C-means and aggregated scoring techniques* (Doctoral dissertation, BRAC University).

[22] Ghosh, P. P., Shahariar, R., & Khan, M. A. H. (2018). A rule based extractive text summarization technique for Bangla news documents. *International Journal of Modern Education and Computer Science*, *11*(12), 44.

[23] Divya, S., & Sripriya, N. (2024). Unified extractive-abstractive summarization: a hybrid approach utilizing BERT and transformer models for enhanced document summarization. *PeerJ Computer Science*, *10*, e2424.

[24] Masri, S., Raddad, Y., Khandaqji, F., Ashqar, H. I., & Elhenawy, M. (2024, July). Transformer models in education: Summarizing science textbooks with AraBART, MT5, AraT5, and mBART. In *International Conference on Intelligent Systems, Blockchain, and Communication Technologies* (pp. 286-300). Cham: Springer Nature Switzerland.

[25] Chalkidis, I., Fergadiotis, M., Malakasiotis, P., Aletras, N., & Androutsopoulos, I. (2020). LEGAL-BERT: The muppets straight out of law school. *arXiv preprint arXiv:2010.02559*.

[26] Masih, S., Hassan, M., Fahad, L. G., & Hassan, B. (2025). Transformer-Based Abstractive Summarization of Legal Texts in Low-Resource Languages. *Electronics*, *14*(12), 2320.

[27] Akhil, S. (2016). An overview of tesseract OCR engine. In *A seminar report. Department of Computer Science and Engineering National Institute of Technology, Calicut Monsoon*.

[28] Emon, M. I. H., Iqbal, K. N., Mehedi, M. H. K., Mahbub, M. J. A., & Rasel, A. A. (2022, August). A review of optical character recognition (ocr) techniques on bengali scripts.

In *International Conference for Emerging Technologies in Computing* (pp. 85-94). Cham: Springer Nature Switzerland.

[29] Mukherjee, S., Tyagi, H., Tyagi, P., Singh, N., & Bhardwaj, S. (2023). OCR using python and its application. Journal of Computer Science, 12(3), 45-58.

[30] Hayat, S. A. I., Das, A., & Hoque, M. M. (2023, December). Abstractive bengali text summarization using transformer-based learning. In *2023 6th International Conference on Electrical Information and Communication Technology (EICT)* (pp. 1-6). IEEE.

[31] Zhang, Y., Jin, H., Meng, D., Wang, J., & Tan, J. (2024). A comprehensive survey on process-oriented automatic text summarization with exploration of llm-based methods. *arXiv preprint arXiv:2403.02901*.

[32] Chhikara, G., Sharma, A., Gurucharan, V., Ghosh, K., & Chakraborty, A. (2024). LaMSUM: A Novel Framework for Extractive Summarization of User Generated Content using LLMs. *rXiv*, *2406*, v1.

[33] Basyal, L., & Sanghvi, M. (2023). Text summarization using large language models: a comparative study of mpt-7b-instruct, falcon-7b-instruct, and openai chat-gpt models. *arXiv preprint arXiv:2310.10449*.

[34] Rony, M. A. T., & Islam, M. S. Evaluating Large Language Models for Summarizing Bangla Texts. In Eighth Widening NLP Workshop (WiNLP 2024) Phase II.

[35] Akter, M., Çano, E., Weber, E., Dobler, D., & Habernal, I. (2025). A Comprehensive Survey on Legal Summarization: Challenges and Future Directions. *arXiv preprint arXiv:2501.17830*.

[36] Wang, J. D., Chang, D., Meng, F. Q., & Qu, G. (2024). A comprehensive survey and prospect of cross-lingual summarization method research. *J. NETWORK INTEL.*, *9*, 384.

[37] Steinberger, J., & Ježek, K. (2009). Evaluation measures for text summarization. *Computing and Informatics*, *28*(2), 251-275.

[38] Aharoni, R., Narayan, S., Maynez, J., Herzig, J., Clark, E., & Lapata, M. (2022). mface: Multilingual summarization with factual consistency evaluation. *arXiv preprint arXiv:2212.10622*.

[39] Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., ... & Zettlemoyer, L. (2019). BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

[40] Rallapalli, S., Gallagher, S., Mellinger, A. O., Ratchford, J., Sinha, A., Brooks, T., ... & Brown, B. (2025). Fine-Tuning LLMs for Report Summarization: Analysis on Supervised and Unsupervised Data. *arXiv preprint arXiv:2503.10676*.

[41] Järvinen, E. (2024). Long-input summarization using large language models.

[42] Akashvarma, M., Yashaswini, G., Keerthana, E., Siddharth, S., Saipooja, K., & Doss, S. A Comprehensive Review of Large Language Models in Abstractive Summarization of News.

[43] Pilault, J., Li, R., Subramanian, S., & Pal, C. (2020, November). On extractive and abstractive neural document summarization with transformer language models. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)* (pp. 9308-9319).

[44] Mullick, A., Bose, S., Saha, R., Bhowmick, A. K., Vempaty, A., Goyal, P., ... & Kokku, R. (2024). Leveraging the power of llms: A fine-tuning approach for high-quality aspect-based summarization. *arXiv preprint arXiv:2408.02584*.

[45] Yuan, H., Hong, S., & Zhang, H. (2025). StrucSum: Graph-Structured Reasoning for Long Document Extractive Summarization with LLMs. *arXiv preprint arXiv:2505.22950*.

[45] Zhang, H., Liu, X., & Zhang, J. (2023). Extractive summarization via chatgpt for faithful summary generation. *arXiv preprint arXiv:2304.04193*.

[46] Favish, A. (2019). Data Capture Automation in the South African Deeds Registry using Optical Character Recognition (OCR).