MARMARA UNIVERSITY

INSTITUTE FOR GRADUATE STUDIES
IN PURE AND APPLIED SCIENCES

Abstractive Legal Text Summarization Using Attention Mechanisms

RAFAH ALOMAR

MASTER THESIS

Department of Computer Science and Engineering

Thesis Supervisor

Assoc. Prof. Dr. Murat Can Ganiz

ISTANBUL, 2024

MARMARA UNIVERSITY

INSTITUTE FOR GRADUATE STUDIES
IN PURE AND APPLIED SCIENCES

Abstractive Legal Text Summarization Using Attention Mechanisms

RAFAH ALOMAR
(524120005)

MASTER THESIS

Department of Computer Science and Engineering

Thesis Supervisor

Assoc. Prof. Dr. Murat Can Ganiz

ISTANBUL, 2024

MARMARA UNIVERSITY

INSTITUTE FOR GRADUATE STUDIES
IN PURE AND APPLIED SCIENCES

Rafah Alomar, a Master of Science student of Marmara University Institute for Graduate Studies in Pure and Applied Sciences, defended her thesis entitled, "**Abstractive Legal Text Summarization Using Attention Mechanisms**", on 13.03.2024 and has been found to be satisfactory by the jury members.

Jury Members

Assoc. Prof. Dr. Murat Can Ganiz (Advisor)
Marmara University, Department of Computer Engineering .....

Assoc. Prof. Mustafa Ağaoğlu (Jury Member)
Marmara University .....

Prof. Dr. Mehmet Fatih Amasyalı (Jury Member)
Yıldız Teknik University .....

APPROVAL

Marmara University Institute for Graduate Studies in Pure and Applied Sciences Executive Committee approves that Rafah Alomar be granted the degree of Master of Science in Department of Computer Engineering, Computer Engineering Program on ..... (Resolution no: .....).

Director of the Institute

ACKNOWLEDGEMENT

I extend my heartfelt gratitude to the many individuals who have contributed to this research. First and foremost, I would like to express my deep appreciation to my advisor, Murat Can Ganiz, for his immense support and guidance throughout this project. His insights and expertise have been invaluable in shaping this study.

I am also grateful to Cihan Erdoğanyılmaz for his valuable insights into the application of AI in the Turkish legal domain. My thanks also go to Mehmet Selman Baysan and Fatih Satı for their significant contributions in collecting Turkish legal data from public resources, which formed the backbone of this study.

Special acknowledgment is owed to Prof. Zafer İçer, whose leadership in organizing and labeling the data has been instrumental. His guidance in standardizing the summarization process and reviewing our results critically has greatly enhanced the quality of this research.

I extend my appreciation to domain experts Elif Tütüncü, Merve Nur Atalay, Beyza Mülayim, Doğa Yıldız, Elifsu Çoban, Sümeyye Çağırıcı, İlayda Demir, Öykü Ecem Şeflek, Aytuğ Şenkal, Gülce Sevim Küçük, Emirhan Akyol, Büşra Nimet Oğuz, Öykü Kır, Rabia Çiçek, Sena Nur Coşgun, Yusuf Zahit Şimşek, Melisa Ayça Laçinel, Bengü Anış Karslı, Önder Semih Adalı, Şule Doğan, İlayda Parlak, and Sevde Nur Kaya for their meticulous efforts in labeling the dataset. Their dedication has been a cornerstone in the success of this research.

I would like to clarify that this research was not funded by any specific grants from public, commercial, or not-for-profit sectors.

Finally, I extend my heartfelt thanks to my family for their unwavering support and encouragement. Your belief in my journey has been instrumental in the completion of this thesis. **03.2024** **RAFAH ALOMAR**

TABLE OF CONTENTS

ABSTRACT

ABSTRACTIVE LEGAL TEXT SUMMARIZATION USING ATTENTION MECHANISMS

Keywords: Abstractive Summarization, Legal Text Summarization, Pre-trained Language Models, Transformers

Automating the summarization of legal documents can save significant time for legal professionals by distilling complex, terminology-heavy texts. In the Turkish legal domain, most existing work focuses on extractive summarization methods. Our study, the first to explore abstractive summarization for Turkish legal documents, compiled a large dataset of higher court decisions and summaries. The training set comprises 13,000 summaries generated using ChatGPT, while the test set contains 2,922 summaries created by Law Faculty students at Marmara University. We experimented with several pretrained transformer models, fine-tuning and evaluating them using our datasets. Although extractive methods outperformed abstractive ones in ROUGE scores, the abstractive approach generated more coherent and concise summaries. In terms of F1 scores, BERT2BERT models excelled, BART achieved the highest precision with a score of 0.44, and GPT-2 yielded the best recall results. This research serves as a foundational step for the future development of abstractive summarization techniques in the context of Turkish legal documents.

ÖZET

ABSTRACTIVE LEGAL TEXT SUMMARIZATION USING ATTENTION MECHANISMS

Anahtar Kelimeler: Hukuk Metni Özetleme, Abstraktif Özetleme

Yasal belgelerin özetlenmesini otomatikleştirmek, karmaşık, terminoloji açısından ağır metinleri damıtarak hukuk uzmanları için önemli miktarda zaman kazandırabilir. Türk hukuk alanında, mevcut çalışmaların çoğu ekstraktif özetleme yöntemlerine odaklanmaktadır. Türkçe hukuk belgeleri için soyut özetlemeyi keşfeden ilk çalışma olan araştırmamız, yüksek mahkeme kararları ve özetleri içeren büyük bir veri seti hazırladı. Eğitim seti, ChatGPT kullanılarak oluşturulan 13.000 özet içerirken, test seti Marmara Üniversitesi Hukuk Fakültesi öğrencileri tarafından oluşturulan 2.922 özet içermektedir. Veri setlerimizi kullanarak birkaç önceden eğitilmiş transformatör modelini denedik, ince ayarladık ve değerlendirdik. Ekstraktif yöntemler ROUGE puanlarında soyut yöntemlerden daha iyi performans gösterse de, soyut yaklaşım daha tutarlı ve özlü özetler oluşturdu. F1 puanları açısından, BERT2BERT modelleri üstünlük gösterdi, BART en yüksek hassasiyeti 0,44 puanla elde etti ve GPT-2 en iyi geri çağırma sonuçlarını verdi. Bu araştırma, Türkçe hukuk belgeleri bağlamında soyut özetleme tekniklerinin gelecekteki gelişimi için temel bir adım oluşturmaktadır.

LIST OF FIGURES

LIST OF TABLES

1 INTRODUCTION

1.1 Background and Motivation

The evolution of Natural Language Processing (NLP) has been heavily influenced by the advent of deep learning, marking a significant paradigm shift in its applications. This section delves into the development of Sequence-to-Sequence (Seq2Seq) and transformer architectures, which have revolutionized NLP research. The introduction of encoder and decoder processes in Seq2Seq models laid the groundwork for sophisticated language processing tasks. Furthermore, the advent of transformer architectures, with their unique parallelization and self-attention mechanisms, opened new avenues in NLP. The integration of pre-trained transformer models like BERT and GPT-2 into Seq2Seq frameworks led to the creation of innovative models such as BERT2BERT and BERT2GPT2. The section also discusses the emergence of transformer-based Seq2Seq models like BART and T5, highlighting their impact on the field.

1.2 Objectives of the Study

This study aims to explore and contrast various NLP architectures, particularly in the context of summarizing legal Turkish texts. A key objective is to compare the effectiveness of these models against decoder-based models, examining their respective advantages and drawbacks. The study also assesses how these models perform compared to extractive summarization methods, such as frequency-based and graph-based strategies. A novel aspect of this research is the introduction of a new dataset, specifically designed for this summarization task. This dataset, comprising 13,000 ChatGPT-labeled text-summary pairs for training and 2,922 human-labeled pairs for testing, is expected to significantly contribute to future research in the domain of abstractive summarization.

1.3 Thesis Structure and Organization

In this thesis, we progress through a series of inter-linked sections, each logically sequenced to build a clear and comprehensive understanding of the technologies at hand. Initially, we lay a foundation of knowledge, gradually moving towards a critical comparison of results. This methodical approach not only facilitates a deep dive into the underlying technologies but also enables a nuanced contrast of experimental scores and observations, ensuring a thorough and informed analysis:

Introduction: This initial segment establishes the foundation of our research. It delineates the overarching theme, lays down the primary objectives, and underscores the significance of the study in the broader context of Natural Language Processing.

Related Work: This section delves into a critical examination of previous research and literature pertinent to our study. This exploration not only contextualizes our research within the existing scholarly discourse but also highlights the niche our work intends to fill.

Methodology: The core of our investigative process is detailed in this section. It outlines the specific methods and approaches employed, offering transparency into the procedural aspects of our research, from model selection to data analysis techniques.

Experiments: This segment is dedicated to the practical application of our methodologies. It chronicles the experimental setup, describes the dataset utilized, and presents a thorough analysis of the findings, enriching the thesis with empirical evidence.

Conclusions: Concluding the thesis, this section synthesizes the key insights gleaned from our research. It reflects on the implications of our findings, discusses the contribution to the field, and proposes potential avenues for future research endeavors."

2 RELATED WORK

2.1 Overview of Summarization Techniques

Summarization techniques are methods for generating a shorter version of a text while preserving the main ideas and key information. There are two main types of summarization techniques: extractive and abstractive.

Extractive summarization works by identifying the most important sentences in the original text and extracting them to form the summary. This is done using a variety of methods:

- Keyword extraction involves identifying the most important words in a text. Techniques like frequency analysis are often used, where frequently occurring words are considered key to the text's meaning.
- Position-Based Extraction assumes that the position of text in a document (e.g., the beginning or end of paragraphs) indicates its importance. Position-based extraction focuses on these areas to summarize content.
- Cue phrase extraction leverages certain phrases indicating importance or summarization points in a text. Phrases like "in conclusion" or "the main point is" often signal key content.
- In graph-based summarization, the text is represented as a graph, with sentences as nodes and relationships between them as edges. Algorithms like PageRank are then used to identify the most important sentences.
- Latent Semantic Analysis (LSA) uncovers the hidden (latent) relationships within the text. It reduces the text to its semantic structure, identifying patterns that help in summarizing the content.
- Lexical Chains involves identifying sequences of related words (lexical chains) in the text. By examining these chains, the most representative sentences can be selected for the summary.
- Topic modeling involves discovering topics prevalent in a text. Techniques like Latent Dirichlet Allocation (LDA) are used to cluster words into topics, which then guide the summarization process.

Abstractive summarization works by generating a new summary text that conveys the

main ideas of the original text in a concise and coherent way. This is a more challenging task than extractive summarization, as it requires understanding the meaning of the text and being able to generate new text that is both accurate and informative.

- Using deep learning, neural networks can generate entirely new sentences to summarize text. They understand the context and semantics to create coherent, concise summaries.
- This method treats summarization as a translation problem. The original text is "translated" into a shorter form, retaining the essential information.
- Statistical methods use mathematical techniques to determine the significance of sentences and generate summaries by combining the most significant ones.
- Template-based methods involve creating a template or structure which the summary will follow. The content is then mapped onto this template.
- Heuristic methods use rules and guidelines, often based on linguistic analysis, to extract and rephrase parts of the text into a summary.

Due to some limitations existing in each approach, many studies and applications combined several methods in what is known as the hybrid summarization methods. One approach is to combine abstractive and extractive methods. This approach first extracts key sentences or fragments and then rephrases them abstractively to produce a coherent summary. Another hybrid approach is to combining statistical methods with machine learning algorithms to improve the accuracy and relevance of summaries. Combining Heuristic and Machine Learning merges heuristic guidelines with machine learning techniques, improving the adaptability and effectiveness of summarization.

Apart from the methodology, other applications of summarization can be differentiated by the source and output of the pipeline such as in the case of Multi-Document Summarization and Cross-Lingual Summarization. Multi-Document Summarization approach involves summarizing information from multiple documents, focusing on extracting and condensing information from various sources into a coherent summary. Cross-lingual summarization involves summarizing text in one language and presenting the summary in another, requiring understanding and translation across languages.

Finally, summarization for Specific Domains involves creating summaries tailored to specific fields like news, science, or law, where domain knowledge is crucial for understanding and summarizing the content effectively.

2.2 Legal Text Summarization Challenges

Legal language summarization is a challenging task due to the unique characteristics of legal documents. These challenges include:

- Complexity of legal language: Legal language is often complex and technical, with specialized terminology and jargon. This can make it difficult for automatic summarization systems to understand the meaning of the text and generate accurate summaries.
- Length of legal documents: Legal documents can be very long and complex, with multiple sections, sub-sections, and paragraphs. This can make it difficult for automatic summarization systems to identify the most important information and generate concise summaries.
- Structure of legal documents: Legal documents have a unique structure, with specific sections, such as the introduction, statement of facts, arguments, and conclusion. This can make it difficult for automatic summarization systems to identify the most important sections and generate summaries that follow the same structure.
- Importance of legal citations: Legal documents often contain citations to other legal documents, statutes, and case law. These citations are important for providing legal support for the arguments made in the document. However, they can also make it difficult for automatic summarization systems to identify the most important information and generate accurate summaries.
- Nuances of legal language: Legal language is often nuanced and subtle, with multiple interpretations of the same text. This can make it difficult for automatic summarization systems to generate summaries that are both accurate and complete.
- Domain-specific knowledge: Legal summarization requires domain-specific knowledge of the law. This can make it difficult for general-purpose automatic summarization systems to generate accurate summaries of legal documents.
- Lack of large-scale training data: There is a lack of large-scale training data for legal summarization. This can make it difficult to train automatic summarization systems to perform well on this task.

- Evolving nature of legal language: Legal language is constantly evolving as new laws are enacted and court decisions are made. This can make it difficult for automatic summarization systems to keep up with the latest changes in legal terminology and jargon.

Additionally, there are challenges that are specific to abstractive summarization of legal documents:

- Preserving legal reasoning: Abstractive summarization systems need to be able to capture the legal reasoning in the original document and generate summaries that accurately reflect the arguments and conclusions of the document.
- Avoiding legal errors: Abstractive summarization systems need to be able to generate summaries that are legally accurate and do not introduce errors or misinterpretations.
- Maintaining consistency with legal style: Abstractive summarization systems need to be able to generate summaries that follow the conventions of legal style and do not introduce any inconsistencies or inaccuracies.

Another major challenge is the scarcity of resources necessary for producing these models. There is a notable deficiency in comprehensive training datasets specifically for summarization in the Turkish legal domain. This shortfall presents significant challenges in training automatic summarization systems to perform effectively in Turkish legal text summarization.

2.3 Abstractive vs. Extractive Summarization

Up until the early 2000s, most automatic text summarization techniques used extractive methods [1]. Popular extractive methods involved scoring sentences using various algorithms, such as term frequency (TF) or inverse term frequency (TF-IDF) [2][3][4], topic modeling [5], latent semantic analysis (LSA) [6], and Bayesian topic models [7][8]. In the extractive approach, sentences are assigned intermediate representations, and based on those representations, a score is computed for each sentence. The top K sentences with the highest scores are then selected and combined to form the summary of the text.

Initial studies in the field of abstractive summarization employed deep learning models, specifically Seq2Seq architectures [9] [10]. Prominent deep learning architectures such as

Convolutional Neural Networks (CNN) [11] and Long Short-Term Memory (LSTM) [12] were tested [13]. Another significant approach applied in this domain is the Generative Adversarial Network (GAN) [14][15]. GAN operates on the premise of two neural networks, a generator and a discriminator, working against each other. The generator creates artificial output (e.g. text), and the discriminator determines whether the output is genuine or artificially generated. Through this competitive process, the generator learns to produce more realistic outputs to fool the discriminator, hence improving the quality of generated summaries over time.

Pointer-Generator [16] systems and Pre-training with Extracted Gap-sentences for Abstractive (PEGASUS) [17] are two prominent approaches designed specifically for abstractive text summarization. Unlike traditional abstractive models that generate summaries purely from scratch, the Pointer-Generator model can decide whether to generate words from the vocabulary or directly copy words from the source text. This copy mechanism enables the model to include important phrases and entities present in the source document, resulting in more accurate and contextually appropriate summaries. With ROUGE-L score of 36.38, the Pointer-Generator system has proven to be effective, especially for handling out-of-vocabulary words and maintaining factual accuracy in the generated summaries. Like Pointer-Generator systems, PEGASUS is also based on the transformer architecture but is trained on two objectives: Masked Language Modeling (MLM) and summarization. PEGASUS is trained using a large corpus of data in an unsupervised manner, allowing it to generate high-quality summaries that capture the essential information from the input text. One of its key strengths is its ability to generalize well across different domains and languages due to its extensive pre-training process. PEGASUS has achieved impressive results, surpassing other pre-trained models across various summarization benchmark datasets, including XSum [18], CNN/DailyMail [9], and Gigaword [19]. It achieved ROUGE-L scores of 39.25, 41.11, and 36.24 on these datasets, respectively.

2.4 Pre-trained Language Models in Summarization

Edunov et al. [20] explored how Embeddings from Language Model (ELMo) [21] augmentation and fine-tuning of pre-trained language model representations could enhance Seq2Seq models. The language model used for summarization task was pre-trained over English newscrawl from the WMT 2018 News dataset (193M sentences, 5B tokens) [22]. In the case of ELMo augmentation, contextualized word embeddings were generated

based on language model representations without making modifications to the underlying language model parameters. On the other hand, fine-tuning involved introducing dropout to the language model output and tweaking learned input word embeddings in the encoder network. By applying these methodologies to the CNN/DailyMail abstractive summarization task, the authors achieved a state-of-the-art ROUGE-L score of 38.47 with the ELMO-based architecture. This research underlined the edge that transfer learning held over pre-training models for embedding applications in various natural language processing tasks. In the legal domain, several works have utilized this technique to generate summaries in different languages by leveraging the advantages of transfer learning over pre-training models in various NLP tasks.

In their research, Shukla et al. [23] implemented pre-trained models like Legal-PEGASUS, BART, and Legal Longformer-Encoder-Decoder (Legal-LED) [24] on specifically curated legal datasets in both English and Indian languages. To create smoother summaries, they utilized a hybrid architecture, where they combined a BERT-based sentence selector with a BART model. Both automated evaluation metrics and expert opinions from legal professionals were utilized to assess the quality of the generated summaries. To handle longer input texts, the researchers experimented with three approaches. Initially, they utilized the Longformer [24] model, specifically designed to accommodate longer token input. Additionally, they explored chunking-based techniques to break down longer input texts into smaller segments, subsequently applying BART or legal-PEGASUS to each chunk. Another method they experimented with for longer documents involved using extractive summarization first, followed by applying abstractive methods to the extracted summary. Their studies reveal that among the pre-trained models, the hybrid and PEGASUS architectures achieved the highest ROUGE-L scores at 0.279 in the Indian extractive dataset. The ROUGE scores for all models experienced a significant boost when fine-tuned, suggesting domain adaptation can lead to better results. For instance, BART's ROUGE-L score increased from 0.271 to 0.404, the hybrid model saw an increase from 0.279 to 0.402, legal-PEGASUS improved from 0.279 to 0.403, and legal-LED jumped from 0.12 to 0.341. Fine-tuned PEGASUS models delivered the highest scores (0.341) on the Indian abstractive datasets however fine-tuned BART models yielded the best results (0.271) on the English abstractive datasets.

In their research study [25], Ouyang et al. introduce an innovative methodology for enhancing the capabilities of language models in task-oriented dialogues. This method

is centered around incorporating human feedback into the training process, guiding the language models to more effectively comprehend and execute instructions. The study outlines the inherent shortcomings of conventional language models. These models are typically geared towards predicting the subsequent word in a sequence based on the preceding words, which often leads to a lack of proficiency in grasping the subtle semantics of natural language. Consequently, these models may falter in accurately following instructions or executing tasks. To mitigate these challenges, the researchers introduce a novel training paradigm, termed the "Instruction Following Loss" (IFL). The IFL is designed to steer the language model towards accurately responding to instructions, factoring in both the specific directive and the context surrounding it. This method is posited to enhance the model's grasp of natural language intricacies, thereby improving its instruction-following capabilities.

The effectiveness of this approach is assessed through various experimental setups. In one such experiment, the language model is trained to follow a series of instructions for constructing a basic object, like a block tower. The model's proficiency in adhering to these instructions is evaluated using a metric known as the "Successful Action Ratio" (SAR). The findings reveal that the IFL-trained model significantly surpasses a traditional model that only focuses on next-word prediction, achieving an SAR of 85% compared to the latter's 63%. Another aspect of the study examines the model's capacity to adapt to unfamiliar instructions, such as assembling a puzzle it has not encountered before. Here again, the IFL-trained model demonstrates superior performance over the baseline model, showcasing its adeptness at generalizing to new scenarios. A critical component of the research is the exploration of the role of human feedback in the training process. The study contrasts the performance of the IFL model with and without human input, revealing a marked improvement in the model trained with human feedback, as evidenced by a 90% SAR compared to 78% for its counterpart. This paper marks a significant stride in advancing how machines understand and react to human instructions, paving the way for more intuitive and effective human-machine interactions. The study's core approach, which employs the Instruction Following Loss (IFL) coupled with human feedback, could potentially be adapted for enhancing the summarization capabilities of language models.

2.5 Summarization in Legal Domain

In [26], court opinions from specific states are collected and cleaned using various techniques like tokenization and lemmatization. In the second step, the text is labeled for its relevance to a human-generated summary using different algorithms such as N-Grams, Longest Common Subsequence (LCS), semantic similarity, and ROUGE score to automatically tag which parts of the court opinion are most relevant for summarization. For extractive summarization, the authors use labeled training data where sentences and paragraphs are marked as relevant or not relevant to the summary. They employ four classifiers: Multinomial Naive Bayes, Decision Trees, Random Forest, and a simple Neural Network with LSTM. These classifiers are used to tag sentences and paragraphs as either relevant or not, which are then assembled to form an extractive summary. In terms of ROUGE-1 scores, the LSTM model excels with a score of 0.55, outperforming other classifiers. Naive Bayes has the lowest performance, registering a score of 0.2. Random Forest and Decision Tree follow closely behind, with scores of 0.5 and 0.48, respectively. For abstractive summarization, a pre-trained PEGASUS language model is fine-tuned for the legal domain. When compared to the $\text{PEGASUS}_{\text{Large}}$ [1]and Legal-PEGASUS models, the fine-tuned version demonstrates superior performance, achieving a ROUGE-1 score of 0.66. $\text{PEGASUS}_{\text{Large}}$ scores significantly lower at 0.39, while the Legal-PEGASUS model registers a score of 0.55.

Huang et al. [27] introduce a two-stage legal judgment summarization model designed to address length-related challenges encountered in pre-trained models. In the initial stage, a convolutional network, BERT and LSTM model are employed to extract crucial sentences and keywords. This involves representing sentences through convolutional and BERT models, utilizing LSTM to annotate sequences of sentence vectors for selection, and extracting keywords using a blend of TF-IDF, Bi-directional Long Short-Term Memory (BiLSTM), and attention mechanisms. Subsequently, the second stage utilizes an abstractive model incorporating Unified Language Model (UniLM) [28] and attention mechanisms to generate summaries. The model's performance was evaluated on the CAIL2020[2] and LCRD[29] datasets. The results showcased the superiority of their approach over various baselines, including Pointer-Generator, BERTSumAbs[30], Keyword Extraction and Summary Generation (KESG) [31], and SummaRuNNer[32]. Notably, their method achieved ROUGE scores of 35.62 and 34.95 for the CAIL2020

[1] `https://huggingface.co/google/pegasus-large`

[2] `https://www.kaggle.com/datasets/weipengfei/sfzy-small`

and LCRD datasets, respectively.

Prabhakar et al. [33] further explore Indian legal summarization by fine-tuning Text-to-Text Transfer Transformer (T5) [34] on a dataset comprised of 350 Indian Supreme Court rulings. Their work focused on the merits of better data pre-processing to get better summaries during the training process. Their pre-processing strategy entailed normalizing several sentence characteristics, such as TF-IDF, noun and verb phrases, proper nouns, cue phrases, and Latin phrases. The normalization procedure involved dividing each feature's value by its respective maximum value. They calculated cosine similarity for each sentence, and the resultant values were then normalized using the maximum cosine similarity value. Moreover, cue phrases and Latin phrases underwent normalization, with the total count of such phrases in the document serving as the normalizing factor. Their results record F-measure scores of 0.65, 0.41, and 0.45 for ROUGE-1, ROUGE-2, and ROUGE-L respectively.

In the domain of Chinese legal summarization, Huang et al. [35] developed an extended model of the Pointer-Generator architecture to include topic information. TRather than user on encoder, their model uses two encoders plus an additional decoder. A bidirectional LSTM (Bi-LSTM) models are used in the encoder to transform the input into hidden states and in the decoder. The hidden states are transformed linearly from the last step and put as input to the decoder. The vanilla encoder-decoder model generates the output sequence based only on this fixed vector representation of the input sequence. However, the proposed model also uses an attention encoder to feed topic word information into the gates of the decoder LSTM. This allows the decoder to generate more relevant and informative output sequences. When tested on a large Chinese legal dataset (total of 12k samples), the model's ROUGE scores outperformed the tried baseline systems. Specifically, the full-length ROUGE F1 evaluation results on the test set show that the proposed model achieves a ROUGE-1 score of 31.96, a ROUGE-2 score of 13.73, and a ROUGE-L score of 30.00. These scores are higher than those of the other models evaluated, including the attention-based sequence-to-sequence model (S2S), pointer-generator network (PTGEN), PTGEN with coverage mechanism (PTGEN-COV), TOPIC-CONVS2S, T-CONVS2S, and LEAD1. The ROUGE metrics are used to evaluate the quality of text summarization systems by comparing the generated summary with a reference summary.

Ingo et al. [36] utilized the Encoder-Decoder architecture to generate abstractive summaries in German court rulings. The previous words are processed using GRU [37],

and the output is used to create an embedding by passing it through two linear layers and ReLU activations. They experimented with altering the attention calculation mechanism based on whether the model is encoding facts or reasonings, calling this method "Guided". Authors also used a variation of this model where instead of using embedding of the previous word the researchers used an embedding of an extracted sentence to actively direct the generation process. They called this variation of the model, "Template". Their results showed that the baseline and guided models perform better than the Template method however abstractive summarization scores remained low compared with the extractive methods.

Yoon et al. [38] presented an abstractive document summarization of Korean legal documents using pre-trained language models BERT2BERT and BART. Their results show that BART models generated higher ROUGE scores than BERT2BERT architecture. BART model showed +3.2 (ROUGE-1), +6.2 (ROUGE-2) and +5.6 (ROUGE-L). While BERT models generated overall good summaries, their sentences had repetitive words and unnatural sentence structure.

In a series of experiments conducted by Robinson et al. [39], several Transformer models such as BART, T5, LED, DISTILBART, and PEGASUS were examined within the legal context. They used two datasets for this purpose - Billsum [40] and GovReport [41], both English. The GovReport dataset, which contains more extensive documents, was deployed for document summarization as well. The evaluation results indicate that the DISTILBART model achieved the highest ROUGE-L score of 51.86 for the Billsum dataset, while the T5 model scored 50.70 for the GovReport dataset. Takale et al. [42] propose to enhance Seq2Seq models with document-context integration, aiming to emulate human comprehension of crucial components. The process involves a three-step methodology for summarizing legal cases, incorporating text reduction strategies like sentence ranking with keywords, dividing input into several chunks, POS tagging followed by selecting the top 15% and bottom 35% of the text, and calculating sentence similarity to identify the upper 40% of sentences. The final step encompasses the application of pre-trained Seq2Seq models, specifically Legal-LED, DISTILBART, and Legal-PEGASUS. The study employs a dataset comprising 45 legal documents from diverse case categories, sourced from *Man upatra* and provided by GNP Legal Law firm, which underwent pre-processing to ensure relevance. Among the models (Legal LED, Legal Pegasus, and DISTILBART), the chunking approach consistently yielded the highest scores. Notably, in terms of ROUGE-L scores, both Legal-LED and Legal-

PEGASUS outperformed DISTILBART across all methodologies, with Legal-LED and Legal-PEGASUS achieving a ROUGE-L of 0.397, and DISTILBART at 0.357.

In the Italian language domain, Dal Pont et al. [43] outline initial findings from the PRODIGIT project which is aimed at leveraging digital technology, specifically AI, to provide support for tax judges and lawyers. The project primarily focuses on generating concise case summaries and extracting pertinent details, such as identifying legal issues, decision-making criteria, and specifying keywords. Various tools and methods for both extractive and abstractive summarization such as IT5 (Italian T5) [44] and GPT models were examined. GPT-4[45] was employed and yielded satisfactory results, as confirmed through evaluations by expert tax judges and lawyers. The project's dataset encompasses 17,000 pairs.

Several strategies have been suggested to tackle the problem of incorrect information or hallucination, which include expanding relevant information, employing re-ranking methods, and fact-checking through POS tagging and entailment relations. In light of these tactics, Feijo et al. introduced a unique model LegalSumm [46] , designed specifically for handling extensive legal documents. This method formulates distinct 'views' of the documents and generates candidate summaries from each segment. To enhance the quality of summarization, LegalSumm incorporates an entailment module to gauge how closely the candidate summaries adhere to the source text. For each input approach, LegalSumm utilizes one entailment model, which is trained using the Hugging Face Transformers library and further fine-tuned with a classification head for the 'fact' or 'fake' category. LegalSumm was tested using the RulingBR dataset [47] (Portuguese) in comparison to internal and external baselines, which showed an improvement in the quality of summarization. In terms of precision, LegalSumm (51 ± 1) outperformed the baseline encoder-decoder Transformer model, and in terms of coherence and coverage, it proved to be superior to BertSumAbs who had a ROUGE-L score of $48(\pm1)$. However, despite these enhancements, the summaries generated were not considered by legal professionals to be particularly factually accurate or an adequate substitute for the original summaries.

2.6 The Turkish Language and NLP

There are a few research papers that concentrate on abstractive summarization in Turkish (general domain), despite the fact that there is no work examining it in legal Turkish.

Baykara et al. [48] conducted experiments with pre-trained transformer models such as BERT for text summarization of news articles in Turkish and Hungarian languages. The authors introduce two datasets, with TR-News being one of them. TR-News consists of 277,573 sample for trainig, 14,610 for validation and 15,379 testing. Different variants of BERT models were used including multilingual and monolingual, cased and uncased, and small and large models. Performance was evaluated with ROUGE metrics (ROUGE-1, ROUGE-2 and ROUGE-L) and compared to the baseline pointer-generator architecture. Their results show that BERT-based models tended to copy from the input document but still produced more novel summaries i.e. more rephrased content, compared to the baseline. These models also produced longer summaries than pointer-generator models however the mBERT-uncased model struggled with words containing Turkish-specific characters. Interestingly, the summaries of the smaller BERT model were better than the larger model in terms of intelligibility and had fewer syntactical errors. They also describe the tokenization methods used and the alterations made to the vocabulary in some models. The methods used for morphological tokenization were SeperateSuffix and CombinedSuffix which are based on the roots and suffixes of the words. Their experiments show that morphological tokenization produced more novel summaries, but can be misleading in cases where the number of out-of-vocabulary (OOV) words is high. On the Turkish dataset presented in the paper, TR-News, the SeperateSuffix model yeilded the best ROUGE-L score at 32.56.

The subsequent research by the same authors [49] advanced their preliminary conclusions by utilizing warm-start strategies in the encoder-decoder architecture using checkpoints (BERT2BERT) from pre-existing models. Their methodology specifically incorporated monolingual T5 (TurkBERT [50]), multilingual BART (mBART) [51], multilingual BERT (mBERT) [52], and multilingual T5 (mt5) [53]. The authors decided to measure the performance of these models across varying data sizes, employing two distinct datasets, TR-News[49] and MLSum [54], along with a combined dataset of the two. The Turkish subset of MLSum is composed of 249,277/11,565/12,775 samples allocated for training, validation, and testing, respectively. Tasks that involved the summarization of Turkish text saw pre-trained encoder-decoder models outdo pointer-generator networks based on Recurrent Neural Networks (RNN). They observed that as the number of training samples increased, the efficacy of the models improved correspondingly. Both the mT5 and BERTurk-cased models demonstrated comparable performance, with mT5 slightly outperforming BERTurk-cased on individual datasets. Specifically, mT5

achieved ROUGE-L scores of 37.96 and 37.60 on the MLSum and TR-News datasets, respectively. On the other hand, BERTurk-cased (in BERT2BERT) obtained ROUGE-L scores of 37.69 and 37.66 on the same datasets. However, when it comes to combined datasets, BERTurk-cased surpassed mT5 significantly, with BERTurk-cased achieving a ROUGE-L score of 39.08 compared to mT5's 38.67. Moreover, the authors found that for both summary and title creation tasks, monolingual BERT models were superior to multilingual BERT models. Issues with encoding in uncased models led to higher average scores for the cased models. Notably, mBART underperformed in comparison to other models on the title creation task, hinting that it might be better suited for challenges requiring more extensive inputs/outputs. The study also evaluated the novelty ratio in summary output for all models across two different tasks. The novelty ratio is the percentage of unique words in the summary that are not present in the source document. BERTurk models (cased and uncased) outperformed both mT5 and mBART models across all datasets. It was suggested that as output length increased, the novelty ratio worsened. This observation was underscored by the significantly higher average novelty ratio for title generation tasks compared to summary generation tasks.

The literature on summarizing Turkish legal texts has not been extensively studied because it is difficult to obtain labeled data. In the only published body of work in text summarization in the legal Turkish language, Gödür [55] examines various approaches for extracting important information from legal text. Using frequency analysis, the first step involves selecting topics and creating summaries. This requires utilizing a 6-word window size to discover related concepts and weighting sentences according to how frequently relevant keywords and concepts appear. Significant cue phrases such as "Sonuç:" indicating the conclusion are discovered to boost the weight of relevant sentences. In this approach, the authors compare the ratio of selected sentences in the generated summary against a reference summary created by an expert. The calculation was performed for 14 examples, resulting in a score of 45.19.

The second approach they use is based on PageRank, which was developed by Erkan and Radev [56] and comprises the creation of a graph of sentence relationships in order to find essential sentences based on their significance in the graph. Like the first method, essential topics are identified using n-gram analysis. Next, each key phrase is associated with the sentence it appears in. After this process is applied to all the sentences, a relationship graph is constructed among the sentences using common key phrases. This creates a graph where sentences related to each other are connected. This step helps

identify related sentences in the text. For the constructed graph, an adjacency matrix is created. This matrix shows the relationships between nodes (sentences) and keeps track of the number of common key phrases between connected nodes (edges). The number of common key phrases between sentences determines the strength of the relationship. Each node is assigned a weight value based on the number of edges it has. These sentences are score and the top sentences are then used to generate a summary. The weighted graph model is an extension of the previously described undirected graph model in which values are assigned to the nodes (sentences) based on the number of connections (edges) they have. However, additionally the values of the connections to each node are assigned based on the position of the sentences in the text. Specifically, the first few sentences and the last few sentences (approximately 10-20% of the sentences in the text) are considered as special nodes. A connection value between 1 and 1.5 is assigned to these special nodes, while all other sentences receive a connection value of 1. In experiments, the best results were attained by assigning priority sentences a value of 1.2. Once again, the authors compare the ratio of selected sentences in the generated summary against a reference summary created by an expert, resulting in an average score of 35.14 for all 14 examples.

The third method is a combination of a topic-based frequency model and a weighted graph model to create a hybrid method. In the proposed method, the top 25 percent of the highest weighted sentences are selected from each method to create two sets of highly scored sentences, set 1 and set 2. From these two sets, a new summary is created by selecting sentences using a predetermined algorithm. This algorithm prioritizes choosing sentences that are present in both set 1 and set 2, and if they are absent from both sets, chooses sentences that are present in either one of the two sets. The final summary, which is limited to 17.5 percent of the original content, is then created by placing the chosen sentences in order of appearance in the original text. The average ratio of selected sentences in the generated summary against a reference summaries created by an expert for this approach had the best value, at 45.62.

2.7 Models Overview in Literature

Table 1: Models Overview in Literature

Language	**Author(s)**	**Year**	**Model(s)**	**Legal Domain**	**Type**
English	Edunov et al.	2019	Seq2Seq with Elmo	No	Abs.
Turkish	Godur, E.	2021	Frequency-based Weighted Graph-based Hybrid	Yes	Ext.
Turkish	Ertem et al.	2022	LSTM-based Seq2Seq	No	Abs.
Turkish, Hungarian	Baykara et al.	2022	Pointer Generator Network, BERT-based Encoder-Decoder	No	Abs.
Turkish	Baykara et al.	2022	Encoder-Decoder using BERT , BART, T5	No	Abs.
Indian, English	Shukla et al.	2022	Legal-Pegasus, BART, Legal-LED, fine-tuning, Hybrid (BERT sentence selector + BART)	Yes	Abs., Ext.
Indian	Prabhakar et al.	2022	T5	Yes	Abs.
Chinese	Huang et al.	2020	Bi-LSTM Encoder + Topic Encoder + Decoder	Yes	Abs.
German	Ingo et al.	2021	GRU	Yes	Abs.
Korean	Yoon et al.	2022	BERT2BERT, BART	Yes	Abs.
English	Robinson et al.	2022	BART, T5, LED, DISTILLBART, PEGASUS	Yes	Abs.
Portuguese	Feiji et al.	2022	LegalSumm	Yes	Abs.
English	Huang et al.	2023	Conv + BERT + LSTM + UniLM	Yes	Abs., Ext.
English	Takale et al.	2023	DISTILLBART, Legal-PEGASUS, Legal-LED	Yes	Abs.

Continued on next page

Table 1 Continued from previous page

Language	Author(s)	Year	Model(s)	Legal Domain	Type
Italian	Dal Pont et al.	2023	T5, GPT	Yes	Abs., Ext.
English	Ghimire et al.	2023	Multinomial Naive Bayes, Decision Trees, Random Forest, LSTM, Legal-PEGASUS, PEGASUS, $\text{PEGASUS}_{\text{CourtOp}}$	Yes	Abs., Ext.

Note: Type refers to the type of summarization investigated. It can either be abstractive (Abs.) or extractive (Ext.).

3 METHODOLOGY

3.1 Dataset Preparation and Preprocessing

3.1.1 Source and Nature of Legal Texts

The Turkish legal domain lacks readily available datasets suitable for summarization tasks. To address this need, a new dataset was constructed using the rulings from the Turkish Court of Cassation, known as "Yargıtay Kararları İçtihat". The Yargıtay serves as the highest court of appeals in Turkey, and its rulings contribute significantly to the development of Turkish jurisprudence. These court decisions not only resolve disputes but also establish legal precedents that guide future decisions on similar cases called "içtihat" in Turkish.

The process involved downloading and curating the Court of Cassation, or CoC in short, rulings. Since these rulings are already available for the public, entity redaction or anonymization is not necessary. Nonetheless, a named entity recognition tool developed in our research lab (@BIGDATA-lab) was employed for improved data privacy by anonymizing remaining sensitive information such as names and addresses,if any. A subset rulings was used for to prepare our dataset, CoCSumTR. These rulings are in free-text form and are not labelled initially. Summaries for 13,000 of these rulings was generated using ChatGPT [25]. This involves submitting the summary to ChatGPT with appropriate prompts, running a length and similarity check using cosine similarity. After which, if the summary satisfies more than 40% similarity check and is not longer than 30% of the original text, the summary is accepted. Another subset of rulings was used for the test dataset. This was initially 100 ruling per annotator and was increased to 300, for more performant annotators. The curated rulings underwent a detailed annotation process wherein law graduate and undergraduate students from Marmara University Law Faculty manually created summaries for each ruling, thereby constructing a supervised dataset. Each data pair in this set comprised a full ruling text and its corresponding summary. A total of 2,922 rulings were annotated by the end of this exercise.

Table 2: Comparison of Dataset Sizes

Dataset	Training	Testing	Language	Domain
CoCSumTR	13,000	2,922	Turkish	Legal
RulingBR	8,497	2,125	Portuguese	Legal
LegalSum	100,000	-	German	Legal
CNN/DailyMail	287,113	11,490	English	News
XSum	204,045	11,334	English	News
MLSUM	1,442,059	128,982	Multilingual	News

3.1.2 Text Length Distribution for Court of Cessation Summarization Dataset (CoCSumTR)

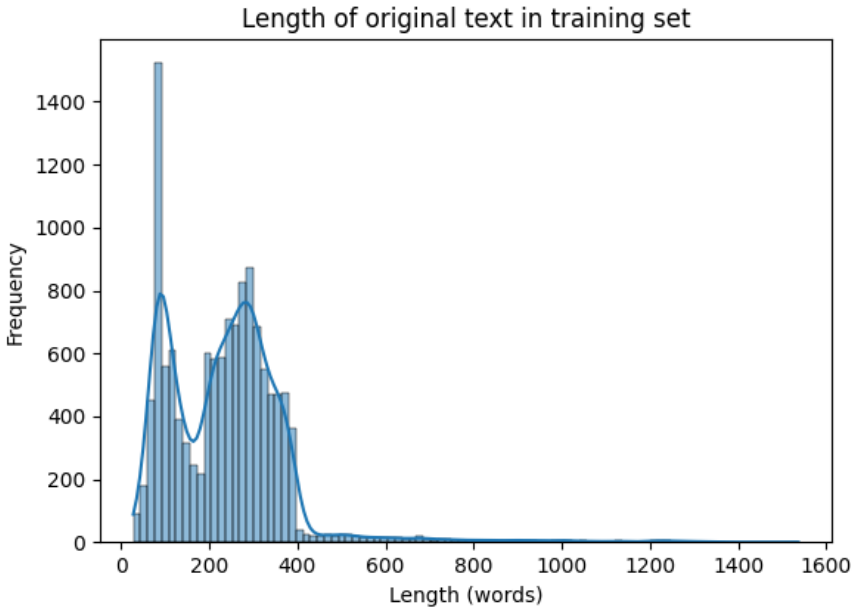

Figure 3.1: Analysis of original text length in training dataset

Figure 3.1 show a historgram of the input text length in the dataset. The X-axis represents the length of original in words and the Y-axis represents frequency. The skewed distribution in the test dataset and the exceptionally high frequency for 80-90 words texts in the training set are attributed to the frequent use of templates in many rulings, where specific details about the reasons and procedures are often omitted, and only the appeal status (rejected, upheld, etc.) is mentioned. Although such rulings were

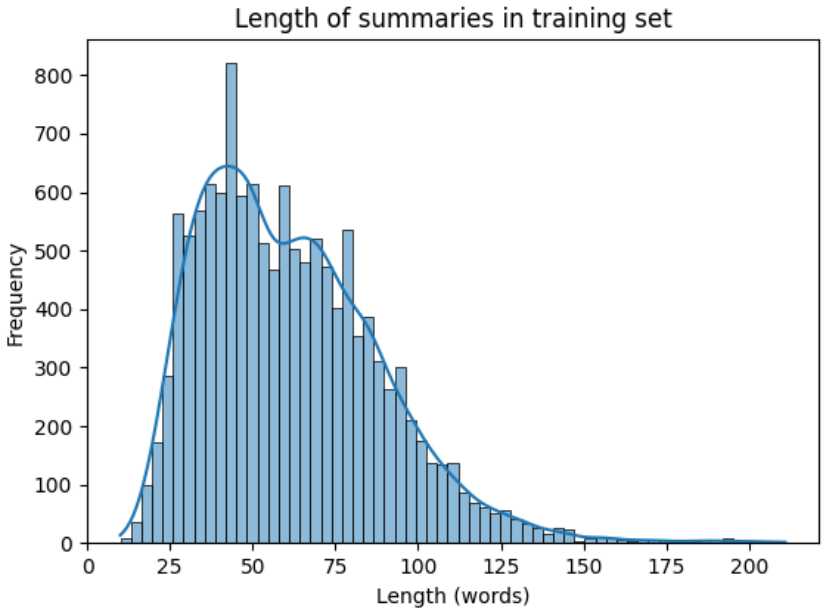

Figure 3.2: Analysis of summary text length in training dataset

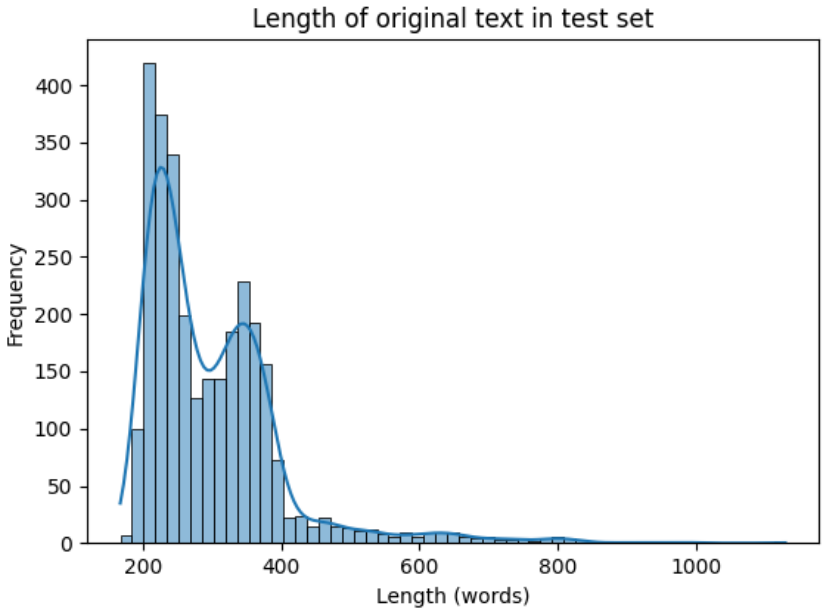

Figure 3.3: Analysis of original text length in test dataset

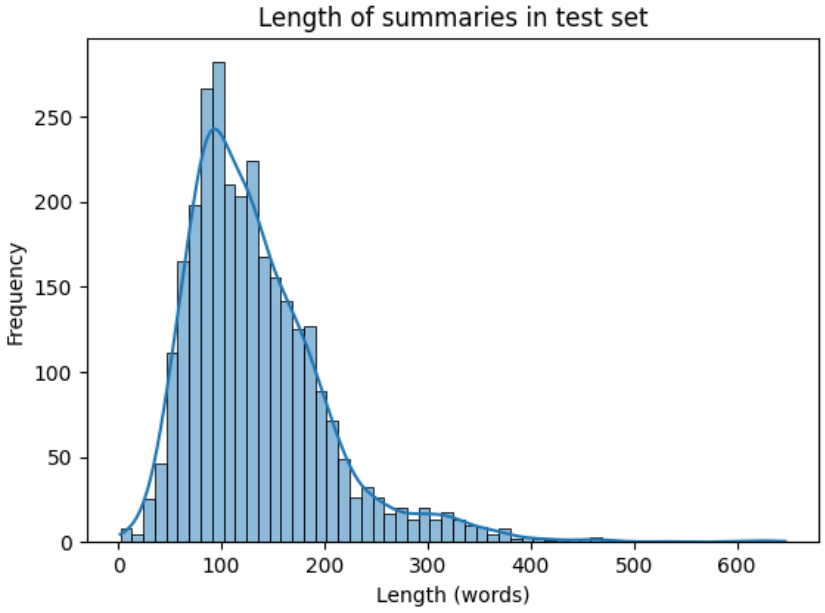

Figure 3.4: Analysis of summary text length in test dataset

avoided in the test data, they were not completely removed. It is possible that during the labeling process, these rulings were favored due to their simplicity and low effort required for labeling. Figures 3.2 and 3.4 show histogram sof the summary length in the training and test datasets. The resulting graph shows a normal distribution which is common in language-related tasks. Summaries need to be long enough to capture key information, but not so long that they become almost as long as the original text. This leads to a natural tendency for summaries to cluster around a certain optimal length, with fewer summaries being significantly shorter or longer. In the ChatGPT-labelling process, this was insured by rejecting summaries longer than 40% of the original text which is also aligned with the guidelines set out for hand-labelling the test dataset.

Table 3: 25^{th} percentile, 75^{th} percentile, mean and standard variation for the length (words) of fields in the CoCSumTR rounded to the nearest whole number

Field	**25%**	**75%**	**Mean**	**Standard Variation**
Train (text)	118	307	237	142
Train (summary)	40	78	61	27
Test (text)	227	348	299	102

Continued on next page

Table 3 – continued from previous page

Field	25%	75%	Mean	Standard Variation
Test (summary)	87	167	134	69

3.1.3 Involvement of Marmara University's Law Faculty Students in Data Labeling

A robust collaborative framework was implemented, leveraging the expertise of both law and computer science scholars. We recruited a group of volunteers comprised of both undergraduate and graduate students from the Faculty of Law at Marmara University. This was accomplished through the assistance of law professors affiliated with Marmara University. Our annotation task was explained to the volunteers in detail, following which they were given access to Doccano[3], our selected tool for sequence-to-sequence labeling. To ensure the volunteers were well-versed with the tool and the task at hand, we organized an initial workshop. This interactive session served as a platform to demonstrate the functionality of Doccano and to provide guidance on text summarization techniques.

To instill a degree of uniformity and standardization in the annotation process, we used a set of reference samples that were previously annotated by law professors. Furthermore, we established different platforms (Email, Google Classrooms, etc.) as a means of ongoing support and discussion. This was intended to field any questions or concerns the annotators might have about the process, including, but not limited to, ideal summarization length, key information capture, and other related topics. While we addressed any technical queries, questions specific to legal language and semantics were fielded by law experts. Bi-weekly meetings were scheduled as a proactive approach to track progress, answer face-to-face inquiries, and ensure the process remained on track. In order to monitor the alignment of student annotations with the provided guidelines, post-processing techniques were utilized. Specifically, we employed cosine similarity and length ratio calculations to assess the consistency of the annotations. These metrics served as a numerical means to evaluate and correct any deviations from the standardized guidelines set forth by the domain experts, thereby maintaining the integrity of our labeled dataset.

[3] `https://doccano.herokuapp.com/`

3.1.4 Unsupervised Dataset for Language Model Fine-Tuning

Alongside the supervised dataset, an unsupervised dataset was also created from another subset of Yargıtay rulings. This dataset consisted of free text from the rulings, with no accompanying summaries. It served a distinct purpose from the supervised dataset, providing raw text for adapting and fine-tuning language models to the legal domain.

Models such as the Turkish GPT-2 and BERTurk were fine-tuned using this dataset. Fine-tuning these pre-existing models allowed us to leverage their language understanding capabilities while adapting them to the specific linguistic nuances and terminology of the legal domain.

The construction of these two distinct yet complementary datasets - one for supervised summarization tasks, and another for unsupervised language model adaptation - represents a substantial contribution to the NLP studies in the Turkish legal domain. Their potential utility extends beyond the scope of the present study, offering valuable resources for future research in this area. We focus on several Seq2Seq architectures that use pre-trained transformers, including BERT2BERT, BERT2GPT2, BART, and T5 with different variations. We present their architecture details and the approach for our experiments. Additionally, we experiment with decoder-based architecture and compare its differences with these models. Since BERT and T5 have a limit of 512 tokens while BART and GPT2 have a limit of 1024 tokens, there are limitations on the input size. However, our dataset inputs may exceed this limit due to their length and the tokenization process, necessitating a more informed approach to handling input cutoff. In legal texts suhc as coc decisions, important information tends to be located at the end. Therefore, we extract the last sentences from the decision while ensuring that the total word count does not exceed 500. To avoid information loss during truncation, we allow for a few extra tokens in case a sentence is still ongoing. Additionally, we set a maximum token limit of 200 for the generated summaries to maintain their conciseness and informativeness.

3.2 Model Selection and Rationale

3.2.1 BERT2BERT BERT2GPT2

BERT2BERT and BERT2GPT2 leverage the strengths of BERT and GPT-2 by using them as encoder and/or decoder components in a Seq2Seq architecture. In the fine-

tuning phase, BERT2BERT and BERT2GPT2 are adapted to the summarization task using our supervised dataset. The encoder ingests the input text and transforms it into a sequence of context-sensitive embeddings. The decoder then takes this encoder output and generates the summary, one token at a time. As the name suggests, BERT2BERT architecture uses pretrained BERT [57] models to initialize both the encoder and decoder components. BERT is designed to pre-train bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers. According to the size of the model, they may have different number of layers, attention heads, and parameters. BERT is originally pre-trained using two unsupervised tasks: masked language model and next sentence prediction [58]. On the other hand, the GPT-2 model specifically uses the decoder of the original transformer model, comprising multiple layers of decoders stacked together. Each block contains a self-attention mechanism and a position-wise feed-forward network, which are each followed by layer normalization and a residual connection. The pretraining objective for GPT-2 is to predict the next token (e.g., word) in a sequence of tokens, given the previous ones. This is why GPT-2 is described as an autoregressive model and requires no additional cross-attention layers to provide language generation.

Decoders in Seq2Seq setups use cross-attention to condition the predictions based on the encoder's output. This is why in both BERT2BERT and BERT2GPT2, cross-attention layers. Cross-attention layer are different from self-attention layers because of the source of its queries, keys and values. In self-attention layers, the queries, keys, and values all come from the same place which is the input sequence. In cross-attention layers, the queries come from one sequence (for instance, the decoder sequence in a transformer), while the keys and values come from another sequence (like the encoder sequence). These additional layers enable the decoder to attend to the encoder's output during each decoding step, a feature absent in the original BERT or GPT-2 models.

We used pre-trained language models on the Turkish language such as Turkish BERT BERTurk[4] [50] and Turkish GPT-2 [5] [59] initilize our architecture. This allowed us to better capture the linguistic nuances, grammar, and specific characteristics of the Turkish language and cut back on training time and computational resources. Additionally, we experiment with some architectures by extending the initial process and training the models on Turkish legal text. Throughout the research, we refer to these extended

[4] `https://huggingface.co/dbmdz/bert-base-turkish-cased`
[5] `https://huggingface.co/redrussianarmy/gpt2-turkish-cased`

models as HukukBERT[6] and HukukGPT2 [7], respectively. For the fine-tuning process, we leverage EncoderDecoderModel from HuggingFace which automatically adds cross-attention layers and initializes the decoder if needed. Once the model has been trained, the model is evaluated on a separate validation dataset that the model has not seen before using ROUGE scores, and the training process is repeated.

3.2.2 MBART

BART [60] is a pre-trained transformer model specifically designed for Seq2Seq tasks. In our research, we leveraged the fine-tuning scripts of the HuggingFace library. Specifically, we utilized the MBartForConditionalGeneration model from HuggingFace's fine-tuning script for BART. BART is a seq2seq model known for its success in various natural language processing tasks. It utilizes an encoder-decoder architecture, incorporating both masked language modeling and denoising auto-encoding objectives during pre-training. We initialized the weights of the architecture with the weights of the pre-trained MBART model (facebook/mbart-large-50 [51]). This is because the original BART model is trained on English corpus whereas MBART was pre-trained on 50 languages including Turkish.

In our fine-tuning process, we imposed a maximum token limit of 200 on the generated summaries to maintain their conciseness. Additionally, it is important to note that the BART models have limitations on the token input size. For example, the input size for the BART model is limited to 1024 tokens. The models are evaluated on a separate validation set, and their performance is measured using ROUGE scores.

3.2.3 mT5

Similarly, for the T5 experiment, we fine-tuned the MT5ForConditionalGeneration model using an MT5 model's (google/mt5-base [53]) pre-trained weights. The MT5ForConditionalGeneration model is based on the T5 (Text-to-Text Transfer Transformer) architecture. T5 is an adaptable model that can be optimized for various NLP tasks, including abstractive summarization. T5 models are trained on a denoising task called Span Corruption. Rather than masking individual tokens like BERT, T5 randomly selects a contiguous span of tokens in the input sequence and replaces it with a noise token. The model is then trained to reconstruct the original sequence, including

[6] `https://huggingface.co/BIGDaTA-Lab/HukukBERT-small`
[7] `https://huggingface.co/BIGDaTA-Lab/hukuk-gpt2-tr`

the corrupted span. A key distinction between T5 and mT5 is that the former undergoes supervised learning during its pre-training stage, which is not the case for the latter. This means that the pre-trained T5 model, even before fine-tuning, has been exposed to various downstream tasks beyond its main unsupervised learning objective and accepts prompt tokens to generate target outputs (e.g. summary). This is not the case for mT5 which requires additional fine-tuning to generate outputs. mT5 has a token limit of 512 and the input text was truncated accordingly.

3.2.4 Turkish GPT2

Decoder-based architectures, such as GPT-2, consist solely of a decoder component and lack an encoder. These models work by predicting the next token in a sequence given the preceding tokens. While they excel at language modeling tasks, their lack of bidirectional context understanding can limit their performance on tasks that require a a thorough comprehension of the original sequence. The underlying fine-tuning process primarily leverages the GPT-2 model, a transformer-based generative language model that utilizes a decoder-only architecture. The fundamental steps of this process include configuring the model, training, and validating the model.

To initialize the fine-tuning process, the model, optimizer, and loss function are set up. A pre-trained Turkish GPT-2 model is used to initialize the weights of the architecture and to benefit from the model's Turkish-specific context knowledge. In this case, the AdamW optimizer is used for training the model, and the learning rate scheduler is configured to adjust the learning rate over the training epochs. The cross entropy Loss function is adopted, which is suitable for multi-class classification problems like text generation. To mitigate the impact of padding tokens during loss computation, an ignore index is specified. The tokens at these indexes will not contribute to the loss. For training, each batch of data is processed in a training loop. The input and label tensors are passed through the model, generating a set of logits. Only the logits and labels after the separator token index are considered for loss calculation, mimicking the functionality of Seq2Seq models. The loss function is applied to these shifted logits and labels.

For consistency with other models, we used beam search as a decoding strategy. The backward pass is computed, and the gradients are clipped to prevent the exploding gradients problem. Afterward, a step of the optimizer is performed, updating the model's weights. The scheduler also takes a step, adjusting the learning rate. This process is

repeated for each batch in the data. The total loss for each batch is accumulated, providing the total loss for an epoch of training.

3.3 Fine-Tuning Strategy for Domain Adaptation

Prior to using in BERT2BERT or BERT2GPT2, we adapt BERTurk and Turkish GPT-2 models to the legal domain by extending their language modelling training over our free text Içtihat dataset. We do this to detect whether understanding of the domain would help generate better-quality summaries. For this, we use the scripts provided by HuggingFace for fine-tuning BERT and GPT-2, respectively. For HukukBERT (BERTurk trained over legal data), using 3 epochs was enough to display a difference in the quality of summaries generated. However, with Hukuk GPT-2 (Turkish GPT-2 trained over legal data), we had to increase the epoch number from 20 to 50 to see the difference in the summarization task results.

3.4 Evaluation Metrics

ROUGE, or Recall-Oriented Understudy for Gisting Evaluation, is a common type of metrics which is used to for evaluation of summarizing. It compares the output of a summary against reference summaries, typically created by humans. The three common variations of ROUGE are ROUGE-1, ROUGE-2, and ROUGE-L.

3.4.1 ROUGE-1

ROUGE-1 measures the overlap of unigram or individual words between the sample generated summary and human-labelled summary. It is calculated using the formula:

$$\text{ROUGE-1} = \frac{\sum_{s \in \{\text{Reference Summaries}\}} \sum_{w \in s} \text{Count}_{\text{match}}(w)}{\sum_{s \in \{\text{Reference Summaries}\}} \sum_{w \in s} \text{Count}(w)}$$

Where:

- $\text{Count}_{\text{match}}(w)$ is the count of each word w in the generated summary that also appears in the reference summary.
- $\text{Count}(w)$ is the count of each word w in the reference summary.

3.4.2 ROUGE-2

ROUGE-2, on the other hand, extends the range to two consecutive words or bi-grams when measure the ratio of overlapping between the summaries. Its formula is:

$$\text{ROUGE-2} = \frac{\sum_{s \in \{\text{Reference Summaries}\}} \sum_{bg \in s} \text{Count}_{\text{match}}(bg)}{\sum_{s \in \{\text{Reference Summaries}\}} \sum_{bg \in s} \text{Count}(bg)}$$

Where:

- $\text{Count}_{\text{match}}(bg)$ is the count of each bigram bg in the generated summary that also appears in the reference summary.
- $\text{Count}(bg)$ is the count of each bigram bg in the reference summary.

3.4.3 ROUGE-L

ROUGE-L focuses on the longest common subsequence (LCS) between the automatically generated and human-generated summary. This means that it considers the similarity on the sentence level by identifying the longest n-grams that co-occur in both summaries. The formula is:

$$\text{ROUGE-L} = \frac{\sum_{s \in \{\text{Reference Summaries}\}} \text{LCS}(s, \text{Generated Summary})}{\sum_{s \in \{\text{Reference Summaries}\}} \text{Length}(s)}$$

Where:

- $\text{LCS}(s, \text{Generated Summary})$ is the length of the longest common subsequence between the reference summary s and the generated summary.
- $\text{Length}(s)$ is the length of the reference summary s.

Each of these ROUGE metrics offers a different perspective on the quality of the summary, with ROUGE-1 and ROUGE-2 focusing on word and phrase accuracy, while ROUGE-L assesses sentence-level fluency and structure.

4 EXPERIMENTAL EVALUATION

4.1 Experimental Setup

For our transformer-based architecture, we chose to select and truncate the terminal sentences so that they did not exceed a length of 500 words. Before settling on this approach, we evaluated alternative strategies such as truncating sentences if they surpassed 500 tokens starting from the beginning of the text, as well as choosing a range of sentences from the beginning, middle, and end without surpassing 500 tokens. However, our experiments led us to conclude that the method of truncating sentences from the end, on average, yielded superior results. Additionally, we consciously avoided interfering with the selection of text input. Our intention was to keep clear distinctions from extractive summary generation approaches, and thus, we refrained from manipulating the input text selection process extensively. Subsequently, we selected the appropriate tokenizer based on the architectural specifications of the model in use. For instance, we employed the T5 tokenizer for T5-based models, the BART tokenizer for BART-based models, and so forth. While employing the BERT2BERT model, we utilized the BERT tokenizer for both encoder and decoder. On the other hand, the BERT2GPT2 model required the application of two distinct tokenizers, BERT tokenizer for the encoder and a GPT-2 tokenizer for the decoder. We did experiments with both cased and uncased variations of the BERTurk however, we eventually settled on continuing with the cased model. This is because the uncased BERT tokenizer strips the accents/dots from the Turkish letters ğ ü ç ö, and caused discrepancies when calculating the ROUGE scores.

Additionally, we conducted experiments using a subset of the dataset consisting of 6,000 samples, to evaluate the performance of these architectures in an under-resourced domain. We ran all our experiments on an Amazon EC2 p3.8xlarge instance, which is equipped with 4 NVIDIA V100 GPUs.

4.2 Model Training

For the BERT2BERT and BERT2GPT2 model variations, we used the default AdamW optimizer as well as learning rate scheduler that adjusts the learning rate during training. The default scheduler is a linear decay scheduler, which we configured to increase linearly during the first 2000 steps of training (warm-up period) and then decay linearly for the remaining steps over the course of training. We implemented early stopping and

a penalty of 2 on output length. Initially, we experimented with a maximum output of 128 tokens for our decoders, however, we found that increasing this limit to 200 led to a higher quality of summaries generated. We also imposed a minimum length of 56 tokens for output to ensure summaries are generated. We trained our BERT2BERT, BERT2GPT2 architecture for 100, T5, BART and GPT-2 architectures for 50 epochs.

4.3 Performance Comparison

4.3.1 Comparison with Extractive Methods

Since there is no previous body of work in abstractive summarization of legal Turkish text, we relied on extractive techniques to draw comparisons with our results. We implemented a range of extractive techniques encompassing graph-based, frequency-based, and clustering-based methodologies. For the first baseline, we implemented a frequency-based extractive summarization algorithm for our dataset. We used a simple approach which works by tokenizing the content into sentences and words, removing Turkish stopwords and punctuation. This is done to focus on the content of the text and to prevent common stops words from being selected in the next step. We then scored each sentence based on the frequency of its terms, and then selected the top 20% of sentences based on their scores to create the summary. For our second baseline, we employed a graph-based text summarization method. First, we tokenized the text into sentences and cleaned them for further processing. Cleaning the data includes removing stop words and new lines and converting to lowercase letters. After cleaning, we utilized a Turkish word2vec model [61] to embed each sentence into a vector space, resulting in a list of sentence vectors. With the sentence vectors at hand, we computed the sentence similarity matrix using cosine similarity, forming the basis for our graph representation. This similarity matrix is then used to create a graph with the help of the NetworkX library, where sentences are nodes and similarity scores are edges. To determine sentence importance, we applied the PageRank algorithm to our sentence graph. The sentences are subsequently ordered based on their respective scores, with the highest-ranking 20% (rounded to the nearest whole number) being chosen for inclusion in the final summary. For our third and final baseline, we use a clustering method proposed by Miller [62] which combines clustering and sentence embeddings for an extractive approach to summarization. It uses BERT to generate sentence embeddings to represent the sentences in the input text, clusters these embeddings and selects representative sentences from each cluster to generate the summary. The number of

clusters is determined dynamically by finding the elbow in the plot of the clustering algorithm's performance versus the number of clusters.

Table 4: Comparison of extractive methods implemented using ROUGE scores with precision, recall, and F1-score for 500 examples

Method	ROUGE1			ROUGE2			ROUGE-L		
	F1	P	R	F1	P	R	F1	P	R
Frequency-based	0.63	0.64	0.72	0.56	0.58	0.63	0.54	0.56	0.62
Bert-based Clustering	0.35	0.56	0.29	0.26	0.42	0.22	0.27	0.44	0.23
Graph-based	0.53	0.63	0.54	0.44	0.54	0.45	0.43	0.53	0.44

4.3.2 Abstractive Model Comparisons

Table 5: Model and Parameter Number

Model	Parameter Number
GPT-2 Summarizer	124,442,880
BERT2BERT	139,608,320
Hukuk BERT2BERT	139,608,320
BERT2GPT2	263,424,000
Hukuk BERT2GPT2	263,424,000
T5 Summarizer	582,401,280
BART Summarizer	610,879,488

The parameter size of a summarization model is a measure of its complexity. A larger model generally has more capacity to learn complex relationships between words and can therefore generate more accurate and fluent summaries. However, larger models also require more training data and can be more computationally expensive to train and run.

In the case of the models listed, BART has the largest number of parameters, followed by T5 and BERT2GPT2. This suggests that these models are the most complex and have the most capacity to learn complex relationships between words. It is also worth noting that BART and T5 models are pre-trained on more diverse tasks that include aspects of both understanding and generating text. For instance, T5 is trained on a

"text-to-text" basis, where almost any NLP task is converted into a text generation problem.

Parameter size is not the only factor that determines the performance of a summarization model, as evident in Table 6. The quality of the training data, the training algorithm, and the hyperparameters can also play a role. While mT5 has a significantly larger number of parameters than BERT2BERT, its performance falls behind. This highlights an important observation: the size of parameters by itself is not consistently a dependable measure of a summarization model's performance. T5, being a multi-task model, is trained on various tasks like machine translation, question answering, etc. Though its large parameter size offers a greater learning capacity, it might not be as optimized for summarization as BERT2BERT, specifically trained for the task. Hukuk BERT2BERT, with the same parameter size as BERT2BERT, performs well due to its specialization for legal texts. Such domain-specific models trained on targeted data often outperform larger generic models like mT5 for specific tasks.

Table 6: Comparison of methods using ROUGE scores with precision, recall, and F1-score for 500 examples

Method	**ROUGE1**			**ROUGE2**			**ROUGE-L**		
	F1	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**
GPT-2 Summarizer	**0.42**	0.49	**0.39**	**0.26**	0.29	**0.24**	0.29	0.33	**0.27**
T5 Summarizer	0.39	0.62	0.31	0.24	0.40	0.19	0.28	0.45	0.25
BART Summarizer	0.38	**0.63**	0.29	0.25	**0.43**	0.19	0.28	**0.48**	0.22
BERT2BERT	0.41	0.62	0.33	**0.26**	0.39	0.22	**0.30**	0.44	0.25
BERT2GPT2	0.40	0.62	0.32	0.25	0.38	0.21	0.29	0.44	0.24
Hukuk BERT2BERT	0.41	0.62	0.34	**0.26**	0.38	0.22	**0.30**	0.44	0.25
Hukuk BERT2GPT2	0.39	0.61	0.32	0.25	0.37	0.21	0.29	0.44	0.24
GPT 3.5	0.34	0.54	0.27	0.17	0.27	0.13	0.22	0.34	0.17

Table 7: Comparison between BERT scores of each model using the BURTTurk

Method	**BERT-score**		
	F1	**P**	**R**
Frequency-based (Ext.)	0.7	0.76	0.72
Bert-based Clustering (Ext.)	0.57	0.5	0.53
Graph-based (Ext.)	0.67	0.66	0.66
GPT-2 Summarizer (Abs.)	0.65	0.57	0.61
T5 Summarizer (Abs.)	0.66	0.57	0.61
BART Summarizer (Abs.)	0.65	0.56	0.6
BERT2BERT (Abs.)	0.66	0.58	0.61
BERT2GPT2 (Abs.)	0.65	0.57	0.61
Hukuk BERT2BERT (Abs.)	0.66	58	0.61
Hukuk BERT2GPT2 (Abs.)	0.65	0.57	0.61
GPT3.5	0.65	0.57	0.61

Table 8: Comparison of methods using average Bleu scores

Method	**BLEU**
Frequency-based (Ext.)	0.44
Bert-based Clustering (Ext.)	0.4
Graph-based (Ext.)	0.44
GPT-2 Summarizer (Abs.)	0.27
T5 Summarizer (Abs.)	0.27
BART Summarizer (Abs.)	0.3
BERT2BERT (Abs.)	0.27
BERT2GPT2 (Abs.)	0.27
Hukuk BERT2BERT (Abs.)	0.27
Hukuk BERT2GPT2 (Abs.)	0.26
GPT3.5	0.22

4.4 Discussion of Results

4.4.1 Interpretation of Findings

From the scores, we can see that extractive methods have higher rouge scores overall than abstractive methods. One explanation for this is that extractive methods generate summaries by selecting sentences from the original text, rather than generating new text which means that the resulting summary is likely to contain a higher percentage of exact matches as the reference summaries, leading to higher ROUGE scores. Another probable reason could be that extractive methods work with a ratio factor or sentence number. In our case, we used the ratio factor since we have variously-sized texts. This means that longer texts would have longer summaries and in turn a better chance of covering more of the original text in the summary. In the abstractive models, we imposed a 200-token output length on the models and penalize lengthier outputs.

Since the length of the summary and the percentage of overlapping words are not good measures of abstractive summaries, we used domain experts to judge the summaries generated by each model. In general, experts expect a good summary to provide practical benefit to lawyers to emphasize which types of crimes the case was opened for, the flow of events in the case and the main elements of the reasoning behind the decision.

In the following example, all summaries contain relevant information about the case, which is about the collection of the unexpropriated immovable property's value. In general, all models are able to correctly emphasize the types of crimes. However, the T5 model's summary seems to be unrelated to the original text and contains information about a retirement case instead.

As expected, extractive approaches generate the most extended summaries because of the ratio parameter, while abstractive summaries are shorter. Summaries generated by models such as the graph-based, GPT-2-based, T5-based, and BART-based suffer from incomplete or cut-off sentences. Apart from degrading the readability of the summary, it also affects how informative the summary is, compared to other methods.

Table 9: Summary generation of different models for sample text

Source	
Original Text	"Taraflar arasındaki kamulaştırmasız el atılan taşınmaz bedelinin tahsili davasından dolayı yapılan yargılama sonunda: Davanın kabulüne ilişkin ilk derece mahkemesinin kararına karşı davalı idare vekilinin istinaf başvurusu üzerine İstanbul Bölge Adliye Mahkemesinin 5. Hukuk Dairesinin istinaf isteminin esastan reddine dair kararı ile birlikte İnegöl 3. Asliye Hukuk Mahkemesinin 2018162 E. - 2018/234 K. sayılı kararının Yargıtay'ca incelenmesi davalı idare vekilince verilen dilekçe ile istenilmiş olmakla, dosyadaki belgeler okunup uyuşmazlık anlaşıldıktan sonra gereği görüşülüp düşünüldü:- K A R A R -**Dava, kamulaştırmasız el atılan taşınmaz bedelinin tahsili istemine ilişkindir.** İlk derece mahkemesince davanın kabulüne ilişkin olarak verilen karara karşı **davalı idare vekili tarafından yapılan istinaf başvurusunun İstanbul Bölge Adliye Mahkemesi 5. Hukuk Dairesince HMK'nın 3531-b-1 inci maddesi gereğince esastan reddine karar verilmiş olup; hüküm davalı idare vekilince temyiz edilmiştir**.Dosyada bulunan kanıt ve belgelere, kararın dayandığı gerekçelere göre; arazi niteliğindeki **İnegöl İlçesi, Boğazköy Mahallesi, 110 ada 62 parsel sayılı taşınmaza 2942 sayılı Kamulaştırma Kanununun 111-f maddesi uyarınca gelir metodu esas alınarak değer biçilmesinde ve alınan rapor uyarınca bedelinin tahsiline ilişkin ilk derece mahkemesince verilen karara karşı yapılan istinaf başvurusunun HMK'nın 3531-b-1 inci maddesi uyarınca esastan reddine karar verilmesinde bir isabetsizlik görülmemiştir.Davalı idare vekilinin temyiz itirazları yerinde olmadığından usul ve yasaya uygun olan hükmün HMK'nun 370. maddesi gereğince ONANMASINA**, davalı idare haçtan muaf olduğundan harç alınmamasına, 15/04/2021 gününde oybirliğiyle karar verildi."

Continued on next page

Table 9 – continued from previous page

Source	
Reference Summary	"**Dava, kamulaştırmasız el atılan taşınmaz bedelinin tahsili istemine ilişkindir.** İlk derece mahkemesince davanın kabulüne ilişkin olarak verilen karara karşı **davalı idare vekili tarafından yapılan istinaf başvurusunun İstanbul Bölge Adliye Mahkemesi 5. Hukuk Dairesince HMK'nın 3531-b-1 inci maddesi gereğince esastan reddine karar verilmiş olup; hüküm davalı idare vekilince temyiz edilmiştir**. Dosyada bulunan kanıt ve belgelere, kararın dayandığı gerekçelere göre; arazi niteliğindeki **İnegöl İlçesi, Boğazköy Mahallesi, 110 ada 62 parsel sayılı taşınmaza 2942 sayılı Kamulaştırma Kanununun 111-f maddesi uyarınca gelir metodu esas alınarak değer biçilmesinde ve alınan rapor uyarınca bedelinin tahsiline ilişkin ilk derece mahkemesince verilen karara karşı yapılan istinaf başvurusunun HMK'nın 3531-b-1 inci maddesi uyarınca esastan reddine karar verilmesinde bir isabetsizlik görülmemiştir.Davalı idare vekilinin temyiz itirazları yerinde olmadığından usul ve yasaya uygun olan hükmün HMK'nun 370. maddesi gereğince ONANMASINA**"

Continued on next page

Table 9 – continued from previous page

Source	
Frequency-based	"Hukuk Dairesince HMK'nın 3531-b-1 inci maddesi gereğince esastan reddine karar verilmiş olup; hüküm davalı idare vekilince temyiz edilmiştir.Dosyada bulunan kanıt ve belgelere, kararın dayandığı gerekçelere göre; arazi niteliğindeki İnegöl İlçesi, Boğazköy Mahallesi, 110 ada 62 parsel sayılı taşınmaza 2942 sayılı Kamulaştırma Kanununun 111-f maddesi uyarınca gelir metodu esas alınarak değer biçilmesinde ve alınan rapor uyarınca bedelinin tahsiline ilişkin ilk derece mahkemesince verilen karara karşı yapılan istinaf başvurusunun HMK'nın 3531-b-1 inci maddesi uyarınca esastan reddine karar verilmesinde bir isabetsizlik görülmemiştir.**Davalı idare vekilinin temyiz itirazları yerinde olmadığından usul ve yasaya uygun olan hükmün HMK'nun 370. maddesi gereğince ONANMASINA,** davalı idare haçtan muaf olduğundan harç alınmamasına, 15042021 gününde oybirliğiyle karar verildi."
Graph-based	"Asliye Hukuk Mahkemesinin 2018162 E. - 2018234 K. sayılı kararının Yargıtay'ca incelenmesi davalı idare vekilince verilen dilekçe ile istenilmiş olmakla, dosyadaki belgeler okunup uyuşmazlık anlaşıldıktan sonra gereği görüşülüp düşünüldü:- K A R A R -Dava, kamulaştırmasız el atılan taşınmaz bedelinin tahsili istemine ilişkindir.İlk derece mahkemesince davanın kabulüne ilişkin olarak verilen karara karşı davalı idare vekili tarafından yapılan istinaf başvurusunun İstanbul Bölge Adliye Mahkemesi 5."

Continued on next page

Table 9 – continued from previous page

Source	
Bert-based (Ext.)	"İlk derece mahkemesince davanın kabulüne ilişkin olarak verilen karara karşı davalı idare vekili tarafından yapılan istinaf başvurusunun İstanbul Bölge Adliye Mahkemesi 5. Hukuk Dairesince HMK'nın 3531-b-1 inci maddesi gereğince esastan reddine karar verilmiş olup; hüküm davalı idare vekilince temyiz edilmiştir. Dosyada bulunan kanıt ve belgelere, kararın dayandığı gerekçelere göre; arazi niteliğindeki İnegöl İlçesi, Boğazköy Mahallesi, 110 ada 62 parsel sayılı taşınmaza 2942 sayılı Kamulaştırma Kanununun 111-f maddesi uyarınca gelir metodu esas alınarak değer biçilmesinde ve alınan rapor uyarınca bedelinin tahsiline ilişkin ilk derece mahkemesince verilen karara karşı yapılan istinaf başvurusunun HMK'nın 3531-b-1 inci maddesi uyarınca esastan reddine karar verilmesinde bir isabetsizlik görülmemiştir."
GPT-2-based	"**İlk derece mahkemesi tarafından yapılan yargılama sonucunda, taraflar arasındaki kamulaştırmasız el atılan taşınmaz bedelinin tahsili davasında, davanın kabulüne ilişkin ilk derece mahkemesinin kararına karşı davalı idare vekilinin yaptığı istinaf başvurusu İstanbul Bölge Adliye Mahkemesinin 5. Hukuk Dairesince esastan reddine karar verilmiştir. Ancak davalı idare vekilinin temyiz itirazları yerinde olmadığından hükmün onanmasına ve harç alınmamasına karar verilmiştir**...u davalı idare haçtan muaf olduğundan harç alınmamasına karar verilmiştir. hallerde, **ilk derece mahkemesince verilen karar HMK'nın 3531-b-1 inci maddesi uyarınca gelir metodu esas alınarak değer biçilmesinde ve bedelin tahsiline ilişkin ilk derece mahkemesince verilen karara karşı yapılan istinaf başvurusunun esas**"

Continued on next page

Table 9 – continued from previous page

Source	
T5-based	"Dava, davacının emeklilik tarihinin tespiti ve emeklilik tarihinin tespit edilmesi istemine ilişkindir. 21. Hukuk Dairesi bozma ilamına uyarak yapılan yargılama sonucunda davanın kısmen kabulüne karar verilmiştir. Taraf vekilleri tarafından temyiz edilen hükmün, dosyadaki yazılara, toplanan delillere ve hükmün dayandığı gerektirici sebeplere göre davacı vekilinin tüm, davalı şirket ile davalı Kurum vekilinin sair temyiz itirazları reddi gerekir. Yargılama giderleri tarafların haklılık oranına göre paylaştırır. Eldeki davada, davacının prim gün ve sigortalılık süresini tamamladığı ancak yaş şartını tamamlamadığından emekliliğe hak kazanamadığı tespit edilmiştir. Bu nedenle, davalı Kurum ve davalı işveren lehine h"
BART-based	"**Bu dava, kamulaştırmasız el atılan taşınmaz bedelinin tahsili istemine ilişkindir. İlk derece mahkemesi davayı kabul etmiş, ancak davalı idare vekili istinaf başvurusu yapmıştır. İstanbul Bölge Adliye Mahkemesi 5. Hukuk Dairesi istinaf isteminin esastan reddine karar vermiş ve hüküm davalı idare vekilince temyiz edilmiştir**. Yargıtay, dosyadaki belgelere ve kararın gerekçelerine dayanarak ilk derece mahkemesinin kararının doğru olduğuna karar vermiştir. **Davalı idare vekilinin temyiz itirazları yerinde olmadığından hükmün onanmasına**"

Continued on next page

Table 9 – continued from previous page

Source	
BERT2BERT	"Bu metin, bir **taşınmaz bedelinin tahsili davası sonunda yapılan yargılama hakkında bir karar metnidir. İlk derece mahkemesi tarafından verilen karara karşı davalı idare vekilinin yaptığı istinaf başvurusu İstanbul Bölge Adliye Mahkemesinin 5. Hukuk Dairesince esastan reddine karar verilmiştir**. Ancak davalı idare vekil tarafından yapılan temyiz itirazları yerinde olmadığından, **hüküm usul ve yasaya uygun olarak onanmıştır**. Davalı idare haçtan muaf olduğu için harç alınmamıştır."
BERT2GPT2	"**Bu metin, bir taşınmaz bedelinin tahsili davası sonunda yapılan yargılama hakkında bir karar metnidir**. İlk derece mahkemesi tarafından verilen karara karşı **davalı idare vekilinin yaptığı istinaf başvurusu İstanbul Bölge Adliye Mahkemesinin 5. Hukuk Dairesince esastan reddine karar verilmiştir**. Ancak **davalı idarenin temyiz itirazları yerinde olmadığından, hüküm usul ve yasaya uygun olarak onanmıştır**. Davalı idare haçtan muaf olduğu için harç alınmamıştır."
Hukuk BERT2BERT	"Bu metin, bir taşınmaz bedelinin tahsili davası sonunda yapılan yargılama hakkında bir karar metnidir. **İlk derece mahkemesi tarafından verilen karara karşı davalı idare vekilinin yaptığı istinaf başvurusu İstanbul Bölge Adliye Mahkemesinin 5. Hukuk Dairesince esastan reddine karar verilmiştir. Ancak davalı idare vekil tarafından yapılan temyiz itirazları yerinde olmadığından, hüküm usul ve yasaya uygun olarak onanmıştır**. Davalı idare haçtan muaf olduğu için harç alınmamıştır."

Continued on next page

Table 9 – continued from previous page

Source	
Hukuk BERT2GPT2	"Bu metin, bir taşınmaz bedelinin tahsili davası sonunda yapılan yargılama hakkında bir karar metnidir. **İlk derece mahkemesi tarafından verilen karara karşı davalı idare vekilinin yaptığı istinaf başvurusu İstanbul Bölge Adliye Mahkemesinin 5. Hukuk Dairesince esastan reddine karar verilmiştir. Ancak davalı idarenin temyiz itirazları yerinde olmadığından, hüküm usul ve yasaya uygun olarak onanmıştır**. Davalı idare haçtan muaf olduğu için harç alınmamıştır."
GPT 3.5	"Taraflar arasında yapılan yargılama sonucunda, kamulaştırmasız el atılan taşınmaz bedelinin tahsili davasının kabul edildiği ve davalı idare vekilinin istinaf başvurusuyla ilgili olarak İstanbul Bölge Adliye Mahkemesi 5. Hukuk Dairesi tarafından istinaf isteminin reddedildiği belirtilmiştir. İnegöl 3. Asliye Hukuk Mahkemesi tarafından verilen kararın Yargıtay tarafından incelenmesi istenmiştir. Dosyadaki belgeler ve uyuşmazlık değerlendirildikten sonra, davalı idare vekilinin temyiz itirazlarının yerinde olmadığı ve hükmün usul"
	End of table

Upon inspection of other examples, we also notice that models, specifically GPT-2-based and BERT2GPT2, suffer from hallucinations. Especially if the original text is dense in terms of numerical data, these models sometimes make up numbers, amounts, or currencies not found in the original text. In Table 10, the reference summary captures the key elements from the text: it discusses an appeal case where the defendant's original sentence of "3 years 4 months" imprisonment was changed to "4 years 8 months 7 days". The appeal reasons by the defendant's lawyer were rejected, and the decision to change the sentence was approved. Ultimately, the file was sent to Çorum 3rd Criminal Court of First Instance and the Chief Public Prosecutor's Office of the Supreme Court in line with Law No. 7165 and Article 304/1 of the Turkish Code of Criminal Procedure (CMK).

The generated summary for BERT2GPT2 shows clear inaccuracies. According to the

original text, the initial decision was given on 16.09.2019 under file number 2019/567 and decision number 2019/2572 by the Regional Court of Justice, not 19.11.2019 with file number 2018/1183 and decision number 2019/2565. Also, the original prison sentence was "3 years 4 months," and it was revised to "4 years 8 months 7 days," but in the generated summary, it's given as "2 years 9 months 10 days." It's also important to mention the defendant was convicted of "intentional wounding," or "kasten yaralama" in Turkish, which the summary does not capture.

For the GPT-2-based model, the summary does not mention one of the key points of the original text where the sentence is increased. There are grammar and syntactical errors as well as incomplete sentences, such as the phrase "sanığın lehine zilyetlikleKararmalarını indiracak şekilde, payına düşen oranında haksız tahrik hükümlerinin". The case numbers, dates, and law article numbers in the summary don't match those given in the original text. The summary also incorrectly identifies the conviction as "nitelikli yağma" (qualified robbery) instead of "intentional wounding".

Table 10: Hallucinations from sample generation from BERT2GPT2

Source	
Original Text	"İstinaf başvurularının kabulü ile yeniden hüküm kurulmak suretiyle sanığın mahkumiyetine dair; ... **Bölge Adliye Mahkemesi 1. Ceza Dairesinin 16.09.2019 tarih, 2019/567 Esas, 2019/2572 Karar sayılı kararı... Bölge Adliye Mahkemesi 1. Ceza Dairesinin 16.09.2019 tarih**, 2019/567 Esas, 2019/2572 Karar sayılı kararının sanık müdafii, katılan vekili tarafından CMK'nin 291. maddesinde belirtilen süre içinde temyiz edildiği anlaşılmıştır.Dosya incelendi.Gereği görüşülüp düşünüldü:İlk derece mahkemesinin 5237 sayılı TCK'nin 86/1, 86/3-e, 87/1-a, d, 29, 62/1 ve 58. maddeleri **uyarınca hükmettiği "3 yıl 4 ay" hapis cezasına dair kararın, bölge adliye mahkemesince kaldırılarak sanığın TCK'nin 86/1, 86/3-e, 87/1-a, d, 29, 62/1 ve 58. maddeleri uyarınca "4 yıl 8 ay 7 gün" hapis cezasına** mahkum edilmiş olması nedeniyle 5271 sayılı CMK'nin 286/2-b maddesi gereğince hükmün temyiz kanun yoluna tabi olduğu belirlenerek yapılan incelemede;... Bölge Adliye Mahkemesi 1. Ceza Dairesinin 16.09.2019 tarih, 2019/567 Esas, 2019/2572 Karar sayılı "istinaf başvurusunun kabulü ile yeniden hüküm kurulmak suretiyle sanığın **kasten yaralama** suçundan mahkumiyetine ilişkin" hükmünün tüm dosya kapsamına göre hukuka uygun olduğu anlaşıldığından; sanıkmüdafii ve katılan vekilinin yerinde görülmeyen **temyiz sebeplerinin reddiyle**, CMK'nin 302/1. maddesi gereğince, isteme uygun olarak **TEMYİZ İSTEMİNİN ESASTAN REDDİ ile HÜKMÜN ONANMASINA**,Dosyanın, 28.02.2019 tarihinde Resmi Gazete'de yayımlanarak yürürlüğe giren 7165 sayılı Kanun'un 8. maddesi ile değişik 5271 sayılı CMK'nin 304/1. maddesi gereğince "Çorum 3. Asliye Ceza Mahkemesine, Yargıtay ilamının bir örneğinin ise ... Bölge Adliye Mahkemesi 1. Ceza Dairesine gönderilmek üzere" Yargıtay Cumhuriyet Başsavcılığına TEVDİİNE, **31.05.2021 gününde oy birliğiyle karar verildi.**"[45]

Continued on next page

Table 10 – continued from previous page

Source	
Reference Summary	"**16.09.2019 tarihli Bölge Adliye Mahkemesi 1. Ceza Dairesi** kararı, sanığın ve katılan vekilinin avukatı tarafından **temyiz** edildi. İlk hapis cezası olan **"3 yıl 4 ay" bölge adliye mahkemesince kaldırıldı ve sanık "4 yıl 8 ay 7 gün" hapis cezasına çarptırıldı**. Yapılan inceleme sonucunda, sanığın kasten yaralama suçundan mahkumiyetine ilişkin hükmün hukuka uygun olduğu belirlendi. Sanık müdafiinin ve katılan vekilinin temyiz talepleri **reddedildi ve hüküm onandı**. Kararın Çorum 3. Asliye Ceza Mahkemesi'ne ve Bölge Adliye Mahkemesi 1. Ceza Dairesi'ne gönderilmesine, 31.05.2021 tarihinde oy birliğiyle karar verildi."
BERT2GP-T2	"Bu metinde, **Bölge Adliye Mahkemesi 1. Ceza Dairesinin 19.11.2019** tarihli 2018/1183 Esas, 2019/2565 Karar sayılı kararının **temyiz** edildiği ve dosyanın incelendiği anlatılmaktadır. İncelemede, sanığın TCK'nin 86/1, 86/3-e, 87/1-d-son, 29, 62. maddeleri uyarımca 2 yıl 9 ay 10 gün hapis cezasına mahkum edilmiş olması nedeniyle 5271 sayılı CMK'nın 286/2-b maddesi gereğince hükmün temyizi kabil kararlardan olduğu belirlenmiştir. Ancak incelemede, sanığa verilen kararın usuli işlemlerinin kanuna uygun olduğu anlaşılmıştır ve katılan vekilinin temyiz itirazları reddedilmiştir."

Continued on next page

Table 10 – continued from previous page

Source	
GPT-2-Based	“ereleyin vegazi veren taraf bir kurum olan limited ve daha fazla iş sözleşmesini sebebiyle niteliklunun ”3 yıl 4 ay’ hapis cezasına mahkumiyetine ilişkin" hükmünü **temyiz** ettiği, Yargıtay Cumhuriyet Başsavcılığı tarafından düzenlenen 15.06.2020 tarihli ve 2020/163 Esas, 2020/2572 Karar sayılı ilamında da belirtildiği üzere, sanığın hükmedilen ceza artırılmış ve açıklandığı gibi, 5271 sayılı CMK’nin 280/1-a maddesi gereğince de ilk derece mahkemelerinden verilen beş yıl veya daha az hapis cezalarını gerektiren nitelikli yağma oluşturup oluşturmayacağı tartışmasızdır.Açıklanmasına karar verilen hükümde ise; Bölge Adliye Mahkemesince sanığın lehine zilyetlikleKararmalarını indiracak şekilde, payına düşen oranında haksız tahrik hükümlerinin”
	End of table

Models generally generated better summaries when the decoder’s maximum output was increased to 200 tokens. In Table 11, we present two summaries generated by two different BERT2BERT models. One model was trained with a maximum output of 128 tokens, and the other with 200. The 128-token summary summary correctly identified the general topic of the text as a guardianship law case. It also accurately described that there were issues of jurisdiction involved in the case However there are inaccuracies in mentioning specific entities such as the "Muratlı Peace Court" and "Ayvalik Public Prosecutor’s Office", which are not in the original text. The role and terms associated with the person under guardianship were also misinterpreted. The 200-token summary effectively communicates the jurisdictional issues and that two different regional courts were involved in examining the files. The notion of a person under guardianship and the restriction on their ability to change the residence is also accurately captured. There are few inaccuracies such as the original text doesn’t refer to the Turkish Civil Code’s Article 411, which is included in the summary. The idea that the guardian was denied the change of residence might also be a slight overinterpretation, as the original text focuses more on the jurisdictional change that would occur with a change of residence.

In Table 12, we see that the summary generated with a 128-token output states that

the case was examined by the Court of Appeals upon the acceptance of the plaintiff's appeal and that the decision was approved. However, the sentence "oy birligi ile 6100 sayılı hmk'nın gecici 3. maddesi geregince karar duzeltme yolu kapalı oldu"(The path to correct the decision was unanimously closed pursuant to the temporary article 3 of the HMK numbered 6100) is not in accordance with the original text. The original text suggests that the route for decision correction remains open for 15 days, not closed. This information is accurately captured in the summary generated by the model with a 200-token output, which also reports the acceptance of the plaintiff's appeal and the approval of the decision by the Court of Appeals. It mentions the 15-day period for decision correction and the date of the decision, aligning more closely with the original text.

Table 11: Summary generation Sample 1 With Max Token Output 128 vs 200 For BERT2BERT

Source	
Original Text	“Kısıtlı hakkında vesayet hukukuna ilişkin olarak açılan davada Yalova Sulh Hukuk ve İstanbul Anadolu 2. Sulh Hukuk Mahkemelerince ayrı ayrı yetkisizlik kararı verilmesi nedeni ile dosyada son karar bölge adliye mahkemelerinin faaliyete geçmesinden sonra verilmiş ise de iki farklı bölge adliye mahkemesinin yargı çevresinde kalan mahkemelerce karşılıklı olarak yetkisizlik kararı verildiğinden ve 5235 sayılı Kanunun 36/3. maddesi gereğince bölge adliye mahkemeleri hukuk dairelerinin görevinin yargı çevresi içerisinde bulunan adlî yargı ilk derece hukuk mahkemeleri arasındaki yetki ve görev uyuşmazlıklarını çözmek olduğundan yargı yerinin belirlenmesi için gönderilen dosya içindeki tüm belgeler incelendi, gereği düşünüldü:- K A R A R -Dosya kapsamında yapılan incelemede Dairemizin 03/11/2020 gün, 2020/7624 Esas, 2020/9111 Karar sayılı ilamı ile;“Dava, kısıtlının taşınmaz mallarının satışı için vasiye izin verilmesine ilişkindir.Yalova Sulh Hukuk Mahkemesi, davacı vekili tarafından kısıtlının adına dava dilekçesinde yazılı taşınmazların satışına izin ve yetki verilmesine dair talep de bulunulduğu, davacı vekili tarafından 2016/1442 esas sayılı dosyada ...’nun kısıtlanmasına dair vesayet davası açıldığı, mahkemece yapılan yargılama sonunda 2016/1442 esas -874 karar ve 01/07/2019 tarihli ek kararıyla yetkisizlik kararı verildiği satışa izin dosyasının da yetkili vesayet makamınca görülmesi gerektiği, kısıtlının yerleşim yeri "... Mah. Anka Çıkmazı sokak ... Blok no ... Daire ... Kadıköy / İstanbul" olduğundan yetkili mahkemenin İstanbul Sulh Hukuk mahkemesi olduğu gerekçesiyle kısıtlının taşınmaz mallarının satışın yetkilivesayet mahkemesinden isteneceğinden davanın yetkisizlik nedeniyle reddine, dosyanın yetkili ve görevli İstanbul Nöbetçi Anadolu Sulh Hukuk Mahkemesine gönderilmesine karar verilmiştir.İstanbul Anadolu 3. Sulh Hukuk Mahkemesi, TMK 462/1 "Taşınmazların alımı, satımı, devredilmesi ve bunlar üzerinde başka bir ayni hak kurulması", TMK 462/14 "vesayet altındaki kişinin yerleşim yerini değiştirmesi" vesayet makamının iznine bağlı olduğundan ve vesayet makamı olan Yalova Sulh hukuk Mahkemesinin 2016/1442 esas sayılı dosyasında TMK 412/1 gereği verilmiş yerleşim yeri değişikliğine izin ve yetkisizlik kararı bulunmadığı ve kısıtlının henüz vesayet makamı değişmemiş olduğu gerekçesiyle mahkemenin yetkisizliğine karar verilmiştir.Türk Medeni Kanununun 411. maddesine göre, “vesayet işlerinde yetki, küçüğün veya kısıtlının yerleşim yerindeki vesayet dairesine aittir. “Aynı Yasanın 19. maddesinde de; “Bir kimsenin ikametgahı, yerleşmek niyetiyle oturduğu yerdir...” hükümlerine yer verilmiştir.TMK’nun 412. maddesinde, vesayet makamının izni olmadıkça vesayet altındaki kişinin yerleşim yerini değiştiremeyeceği belirtildiğinden, bu durumda, vesayet makamı olanYalova Sulh Hukuk Mahkemesinin 04.07.2017 tarih ve 2016/1442 E. - 2017/874 K. sayılı ilamı kararı ile kısıtlanmasına karar verilen ...’nun ...’ın vasisi olmasına karar verildiği,Yalova Sulh Hukuk Mahkemesi tarafından kısıtlıların ikametgahının değiştirilmesine izin verilmediğine göre uyuşmazlığın Yalova Sulh Hukuk Mahkemsinde görülüp çözümlenmesi gerekmektedir.” gereğine değinilerek kısıtlının taşınmaz mallarının satışı için vasiye izin verilmesi dosyasında Yalova Sulh Hukuk Mahkemesinin yargı yeri olarak belirlenmesine karar verildiği anlaşılmaktadır.Dava, satışa izin dosyasında vasinin hissedar olmasından dolayı menfaat çatışması olabileceğinden yeni kayyım atanması istemine ilişkindir.Yalova Sulh Hukuk Mahkemesi, davacı vekili tarafından Yalova Sulh Hukuk Mahkemesinin 2018/952 Esas sayılı dosyasında kısıtlıya ait taşınmazın satışının yapılması için kayyım tayin edilmesi yönünde karar verildiğini, kayyım tayin edilmesi yönünde verilen karar gereği, kendilerine kayyın ettirilmesi amacıyla gerekli izin ve yetki belgesinin 11/02/2019 tarihinde verildiğini belirterek kayyım atanmasını talep ettiği, kısıtlının uyap sisteminden alınan nüfus kayıtlarının tetkikinde, adresinin adresinin "... Mah. ... Çıkmazı Sk. No:...İç Kapı No:.. /İSTANBUL" olduğu, kayyım adayının annesinin İstanbul’da yaşadığını beyan ettiği gerekçesiyle yetkisizlik kararı verilmiştir.İstanbul Anadolu 2. Sulh Hukuk Mahkemesi ise, kendisine kayyım tayini talep edilen kısıtlının yerleşim yeri adresi değiştiğinden bahisle yetkisizlik kararı verilmiş ise de İstanbul Anadolu 21. Sulh Hukuk Mahkemesinin 2020/120 esas sayılı dosyasında karşı yetkisizlik kararı nedeniyle merci tayini yapıldığı ve vesayet dosyasının Yalova Sulh Hukuk Mahkemesinin 2020/987 esas sayılı dosyasında devam ettiği, dava tarihi itibariyle yetkili mahkemenin tayin edileceği, dava tarihi itibariyle ve halen kısıtlının Yalova Sulh Hukuk Mahkemesi yetki çevresinde ve bu mahkemenin vesayeti altında bulunduğu, kısıtlıya kayyım tayin edilmesi yönünden vesayet makamının bulunduğu yerdeki sulh hukuk mahkemesi yetkili olduğu, buna göre de Yalova Sulh Hukuk Mahkemesinin kesin yetkili mahkeme olduğu gerekçesiyle yetkisizlik kararı verilmiştir.Türk Medenî Kanununun 426. maddesinde "Vesayet makamı, aşağıda yazılı olan veya kanunda gösterilen diğer hallerde ilgilisinin isteği üzerine veya re’sen temsil kayyımı atar. Bir işte yasal temsilcinin menfaati ile küçüğün veya kısıtlının menfaati çatışıyorsa..." hükmü; 431. maddesinde ise "Vasinin atanması usûlüne ilişkin kurallar, kayyım ve yasal danışmanın atanmasında da uygulanır." hükmü getirilmiştir.Diğer yandan, aynı Kanunun 411. maddesine göre de, "Vesayet işlerinde yetki, küçüğün veya kısıtlının yerleşim yerindeki vesayet dairelerine aittir." “Aynı Kanunun 19. maddesinde de; “Bir kimsenin ikametgahı, yerleşmek niyetiyle oturduğu yerdir...” hükümlerine yer verilmiştir.TMK’nın 412. maddesinde, vesayet makamının izni olmadıkça vesayet altındaki kişinin yerleşim yerini değiştiremeyeceği hükümlerine yer verilmiştir.Dosya kapsamından, Yalova Sulh Hukuk Mahkemesinin 04/07/2017 tarih ve 2016/1442 E. - 2017/874 K. sayılı ilamı ile kısıtlanarak, ...’ın vasi olarak atandığı, vesayet makamı olan Yalova Sulh Hukuk Mahkemesi tarafından kısıtlıların ikametgahının değiştirilmesine izin verilmediğine göre uyuşmazlığın Yalova Sulh Hukuk Mahkemesinde görülüp çözümlenmesi gerekmektedir.Yukarıda açıklanan nedenlerle, 6100 sayılı HMK’nın 21 ve 22. maddeleri ve 5235 sayılı Kanunun 36/3. maddesi gereğince Yalova Sulh Hukuk Mahkemesinin YARGI YERİ OLARAK BELİRLENMESİNE, 31/05/2021 gününde oybirliğiyle karar verildi.”	

Continued on next page

Table 11 – continued from previous page

Source	
Reference Summary	"**Bu metin, Yalova Sulh Hukuk Mahkemesi ve İstanbul Anadolu 2. Sulh Hukuk Mahkemesi arasında bir yetki uyuşmazlığına ilişkin bir hukuk davasını anlatmaktadır**. Davanın konusu, kısıtlı bir kişinin taşınmaz mallarının satışı için vasiye izin verilmesi ve kayyım atanması istemidir. İki farklı mahkeme ayrı ayrı **yetkisizlik kararı vermiştir**. Yalova Sulh Hukuk Mahkemesi, davanın İstanbul Sulh Hukuk Mahkemesi yetkisinde olduğunu, İstanbul Anadolu 2. Sulh Hukuk Mahkemesi ise Yalova Sulh Hukuk Mahkemesi'nin yetkili olduğunu belirtmiştir. Ancak, incelenen tüm belgeler sonucunda, vesayet makamının bulunduğu Yalova Sulh Hukuk Mahkemesi'nin yargı yeri olarak belirlenmesine karar verilmiştir. Bunun nedeni, kısıtlının ikametgahının d**eğiştirilmesine Yalova Sulh Hukuk Mahkemesi tarafından izin verilmemesi ve kısıtlının vasi olarak atanmasının bu mahkeme tarafından gerçekleştirilmiş olmasıdır.**"
BERT2BERT (128)	"bu metin, bir vesayet hukuku davası hakkında bilgi veriyor. muratlı sulh hukuk mahkemesi ve ele gecen tasınmazların kadastro tespiti icin acılmıstır. ancak, davalı vekili tarafından **yetkisizlik kararı verilmistir**. dosya incelendiginde, kısıtlı adayın yerlesim yeri bakımından ayvalık cumhuriyet bassavcılıgına tevdi edilmistir. istanbul anadolu 2. sulh hukuk mahkemeleri de aynı kararı vermistir. kısıtlı kısıtlanmıs ve vasi olarak atanmıstır. vesayet makamının izni olmadan kısıtlı yerlesim yerini degistiremez. yerlesim yerinin degismesi halinde yetki yeni vesayet dairelerine gecer."

Continued on next page

Table 11 – continued from previous page

Source	
BERT2BERT (200)	Türk Medeni Kanunu'nun 411. maddesi gereğince, vesayet işlerinde yetki küçük veya kısıtlının yerleşim yerindeki vesayet dairesine aittir. **Kısıtlının taşınmazın satışı için yetkisizlik kararı verilmiş ve Yalova Sulh Hukuk Mahkemesi tarafından vesayet altına alınmıştı**. Ancak, iki farklı bölge adliye mahkemesi yargı yeri olarak belirlenmiş ve dosyalar incelendi. Sonuç olarak, kayyımın yerleşim yeri değişikliğine izin verilmemiş ve vesayet dosyasının davalıya ait olması gerektiği belirtilmiştir."
	End of table

Table 12: Summary generation Sample 2 With Max Token Output 128 vs 200 For BERT2BERT

Source	
Original Text	"Taraflar arasındaki kira sözleşmesinden kaynaklanan alacak davasının mahkemece yapılan yargılaması sonucunda, davanın kısmen kabulüne yönelik olarak verilen hükmün, süresi içinde davacı vekili tarafından temyiz edilmesi üzerine; temyiz dilekçesinin kabulüne karar verildikten sonra, dosya içerisindeki bütün kağıtlar okunup gereği düşünüldü:Y A R G I T A Y K A R A R IDosyadaki yazılara, kararın bozmaya uygun olmasına ve delillerin takdirinde bir isabetsizlik bulunmamasına göre, yerinde görülmeyen bütün temyiz itirazlarının reddi ile usul ve kanuna uygun olan hükmün **ONANMASINA**, 6100 sayılı HMK'nın geçici madde 3 atfıyla 1086 sayılı HUMK'nın 440.maddesi gereğince kararın tebliğinden **itibaren 15 günlük süre içerisinde karar düzeltme yolu açık olmak üzere, 31032021 tarihinde oy birliği ile karar verildi.**"
	Continued on next page

Table 12 – continued from previous page

Source	
Reference Summary	"Yargıtay, dosyadaki yazılara ve delillere dayanarak davanın kabulüne yönelik olarak verilen hükmün onanmasına karar verdi. Karar düzeltme yolu, tebliğinden itibaren **15 günlük süre içerisinde açık olmak üzere 31032021 tarihinde oy** birliği ile karar verildi."
BERT2BERT (128 Max token)	"yargıtay, davacı vekilinin temyiz dilekcesinin kabulu uzerine dosya icerisindeki butun kagıtlar incelendi ve usul ve kanuna uygun olan hukmun **onanmasına** karar verildi. kararın duzeltme yolu **15 gunluk sure icerisinde acıktır**. oy birligi ile 6100 sayılı hmk'nın gecici 3. maddesi geregince karar duzeltma yolu kapalı oldu."
BERT2BERT (200 Max token)	"yargıtay, davacı vekilinin temyiz dilekcesinin kabulu uzerine dosya icerisindeki butun kagıtlar incelendi ve usul ve kanuna uygun olan hukmun **onanmasına** karar verildi. kararın duzeltme yolu 15 gunluk sure icerisinde acıktır. **oy birligi ile 31 03 2021 tarihinde verildi.**"
	End of table

4.4.2 Model Strengths and Weaknesses

Table 13: Comparison of weaknesses and strengths of each model

Method	**Strengths**	**Weaknesses**
Extractive Methods	Higher ROUGE scores; longer summaries with more detail.	Lack of novel text generation.
Abstractive Methods	Generation of new text; expert approval for practicality in legal contexts.	Lower ROUGE scores; output limitations; incomplete or cut-off sentences.
BART	Highest ROUGE score among abstractive models.	Incomplete sentences affecting quality.
	Continued on next page	

Table 13 – continued from previous page

Method	Strengths	Weaknesses
T5	Good at emphasizing key types of crimes.	Unrelated content generation; misses crucial information due to syntax or length constraints.
BERT2BERT Variants	Captures summaries effectively.	May miss crucial reasoning information to maintain sentence syntax and length.
BERT2GPT2 Variants	Capable of generating novel text.	Hallucinations issues due to GPT2 in the decoder.
GPT-2	Generation of novel text.	content relevance and completeness issues.
	End of table	

4.4.3 Implications for Legal Practice

- **Balance Between Accuracy and Coherence:** Legal language summarization demands a delicate balance. While extractive methods score higher in terms of accuracy (as measured by ROUGE scores), abstractive methods are better at generating coherent and fluent summaries. However, they sometimes struggle with maintaining the accuracy and relevance of the content, particularly in legal contexts where every detail can be crucial.
- **Importance of Domain-Specific Adaptation:** Legal texts often contain complex structures and specialized terminology. This necessitates models like BERT2BERT, which, while capturing summaries effectively, may miss crucial reasoning details. Tailoring models to better understand and reproduce legal reasoning and jargon is essential.
- **Challenge of Maintaining Contextual Integrity:** Models like T5 and BERT2GPT2 illustrate the challenges in maintaining contextual integrity. T5 sometimes generates content unrelated to the main legal arguments, and BERT2GPT2 has hallucination issues. Ensuring that summaries stay true to the original context and content is critical in legal summarization.

- **Need for Expert Validation:** The use of domain experts to evaluate the practicality of summaries underscores the importance of human judgment in assessing the quality of legal summaries. Automated metrics like ROUGE scores are not always sufficient to capture the nuances required in legal documents.
- **Potential for Complementary Use of Models:** Because each model has its own strength points as well as limitations, a blend of models could give more effective results. For instance, combining extractive methods for accurately capturing crucial details with abstractive methods for coherent and concise rewriting could enhance the overall quality of legal document summaries.
- **Consideration of Summary Length and Detail:** The length and level of detail in summaries are significant factors. Extractive methods tend to produce longer summaries due to their ratio factor, which can be beneficial in covering more content but might also lead to verbosity. In contrast, abstractive models, with their fixed output lengths, may create more concise summaries but risk omitting essential information.

5 CONCLUSIONS AND FUTURE WORK

5.1 Summary of Findings

Our study conducted a thorough analysis of various NLP architectures, focusing on their application in summarizing legal Turkish texts. The findings revealed that transformer-based Seq2Seq models, such as BART and T5, have significantly advanced the field of NLP, especially in abstractive summarization. These models outperform traditional decoder-based models in terms of generating more coherent and contextually relevant summaries. However, they sometimes struggle with accuracy and maintaining the integrity of complex legal reasoning. Extractive methods, while less advanced in terms of linguistic generation, showed higher accuracy in terms of ROUGE scores, yet lacked the ability to reformulate and condense text effectively.

5.2 Contributions

This study makes several key contributions to the field of NLP and legal text summarization:

5.2.1 Novel Datasets

We introduced a novel dataset comprising 13,000 ChatGPT-labeled text-summary pairs for training and 2,922 human-labeled pairs for testing. This dataset is specifically tailored for the task of summarizing legal Turkish texts and is expected to be a valuable resource for future research in abstractive summarization.

5.2.2 Contribution in Turkish Legal Summarization

Our work contributes to the understanding of how different NLP models perform in the context of Turkish legal text summarization. By comparing various architectures, we provide insights into their applicability and effectiveness in a domain that has unique linguistic and structural challenges. This study paves the way for more targeted and effective summarization approaches in legal contexts.

5.3 Limitations

This study, while comprehensive in its scope, encounters several limitations that must be acknowledged:

Inadequacy of ROUGE Scores: The use of ROUGE scores as the primary metric for evaluation may not effectively capture the quality of abstractive summarization. ROUGE focuses on overlap with reference texts, which may not fully represent the summarization capabilities of abstractive models.

Bias Towards Extractive Methods: The test dataset, being extractively labeled, inherently favors models that are more aligned with extractive summarization techniques. This introduces a bias in evaluating performance, potentially undervaluing the nuances of abstractive methods.

Limitations in Handling Longer Documents: The current analysis focuses on models with a maximum token limitation. This constraint means that the study does not fully address the challenges associated with summarizing longer documents, which may require different approaches or models.

Generalization Concerns: The findings, while significant for Turkish legal texts, may not be directly applicable to other languages or legal systems, limiting the generalizability of the results.

5.4 Future Work

Considering the limitations and the evolving landscape of NLP, several avenues for future work are proposed:

Exploring Models for Longer Documents: Future research should focus on models like Reformer and Longformer Encoder Decoder(LED) that are designed to handle longer sequences. This will address the current gap in summarizing extensive legal documents.

Cross-Language Model Adaptation: Fine-tuning models that have shown success in English legal summarization and then transferring this learning to the Turkish legal domain can offer novel insights and enhance model performance across languages.

Investigating Deviations in Model Focus: A deeper analysis into why models like BERT2BERT, BERT2GPT2, and T5 tend to deviate from crucial areas of information in legal texts is necessary. Understanding these tendencies will be crucial in developing

more accurate and reliable summarization models.

Enhanced Evaluation Metrics: Developing or adopting more nuanced evaluation metrics beyond ROUGE scores will be essential. These metrics should better capture the effectiveness of abstractive summarization in preserving crucial legal information and context.

Diverse Dataset Development: Expanding the dataset to include a wider range of legal texts, possibly from different legal systems and languages, will help in refining the models and making the findings more generalizable.

6 REFERENCES

[1] Divakar Yadav, Jalpa Desai, and Arun Kumar Yadav. *Automatic Text Summarization Methods: A Comprehensive Review*. 2022. arXiv: 2204.01849 [cs.CL].

[2] Chin-Yew Lin and Eduard Hovy. "Automated Multi-Document Summarization in NeATS". In: *Proceedings of the Second International Conference on Human Language Technology Research*. HLT '02. San Diego, California: Morgan Kaufmann Publishers Inc., 2002, pp. 59–62.

[3] H. P. Luhn. "The Automatic Creation of Literature Abstracts". In: *IBM Journal of Research and Development* 2.2 (1958), pp. 159–165. DOI: 10.1147/rd.22.0159.

[4] Ani Nenkova, Lucy Vanderwende, and Lucy Vanderwende. "The Impact of Frequency on Summarization". In: 2005. URL: https://api.semanticscholar.org/CorpusID:14102322.

[5] Robert F. Lorch et al. "Effects of Headings on Text Summarization". In: *Contemporary Educational Psychology* 26.2 (2001), pp. 171–191. ISSN: 0361-476X. DOI: https://doi.org/10.1006/ceps.1999.1037. URL: https://www.sciencedirect.com/science/article/pii/S0361476X99910378.

[6] Yihong Gong and Xin Liu. "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 Orleans, Louisiana, USA: Association for Computing Machinery, 2001, pp. 19–25. ISBN: 1581133316. DOI: 10.1145/383952.383955. URL: https://doi.org/10.1145/383952.383955.

[7] Dingding Wang et al. "Multi-Document Summarization using Sentence-based Topic Models". In: *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*. Suntec, Singapore: Association for Computational Linguistics, Aug. 2009, pp. 297–300. URL: https://aclanthology.org/P09-2075.

[8] Feifei Wang et al. "Bayesian Text Classification and Summarization via A Class-Specified Topic Model". In: *Journal of Machine Learning Research* 22.89 (2021), pp. 1–48. URL: http://jmlr.org/papers/v22/18-332.html.

[9] Ramesh Nallapati et al. "Abstractive Text Summarization using Sequence-to-sequence RNNs and Beyond". In: *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*. Berlin, Germany: Association for

Computational Linguistics, Aug. 2016, pp. 280–290. DOI: 10.18653/v1/K16-1028. URL: https://aclanthology.org/K16-1028.

[10] Piji Li et al. "Deep Recurrent Generative Decoder for Abstractive Text Summarization". In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics, Sept. 2017, pp. 2091–2100. DOI: 10.18653/v1/D17-1222. URL: https://aclanthology.org/D17-1222.

[11] Keiron O'Shea and Ryan Nash. "An Introduction to Convolutional Neural Networks". In: *CoRR* abs/1511.08458 (2015). arXiv: 1511.08458. URL: http://arxiv.org/abs/1511.08458.

[12] Sepp Hochreiter and Jürgen Schmidhuber. "Long Short-Term Memory". In: *Neural Comput.* 9.8 (Nov. 1997), pp. 1735–1780. ISSN: 0899-7667. DOI: 10.1162/neco.1997.9.8.1735. URL: https://doi.org/10.1162/neco.1997.9.8.1735.

[13] Panagiotis Kouris, Georgios Alexandridis, and Andreas Stafylopatis. "Abstractive Text Summarization Based on Deep Learning and Semantic Content Generalization". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, July 2019, pp. 5082–5092. DOI: 10.18653/v1/P19-1501. URL: https://aclanthology.org/P19-1501.

[14] Min Yang et al. "Plausibility-promoting generative adversarial network for abstractive text summarization with multi-task constraint". In: *Information Sciences* 521 (2020), pp. 46–61. ISSN: 0020-0255. DOI: https://doi.org/10.1016/j.ins.2020.02.040. URL: https://www.sciencedirect.com/science/article/pii/S0020025520301225.

[15] Min Yang et al. "Hierarchical Human-Like Deep Neural Networks for Abstractive Text Summarization". eng. In: *IEEE Transactions on Neural Networks and Learning Systems* 32.6 (2021), pp. 2744–2757. ISSN: 2162-2388. DOI: 10.1109/TNNLS.2020.3008037. URL: https://doi.org/10.1109/TNNLS.2020.3008037.

[16] Abigail See, Peter J. Liu, and Christopher D. Manning. "Get To The Point: Summarization with Pointer-Generator Networks". In: *CoRR* abs/1704.04368 (2017). arXiv: 1704.04368. URL: http://arxiv.org/abs/1704.04368.

[17] Jingqing Zhang et al. "PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization". In: *CoRR* abs/1912.08777 (2019). arXiv: 1912.08777. URL: http://arxiv.org/abs/1912.08777.

[18] Shashi Narayan, Shay B. Cohen, and Mirella Lapata. "Don't Give Me the Details, Just the Summary! Topic-Aware Convolutional Neural Networks for Extreme Summarization". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, Oct. 2018, pp. 1797–1807. DOI: 10.18653/v1/D18-1206. URL: https://aclanthology.org/D18-1206.

[19] David Graff et al. "English gigaword". In: *Linguistic Data Consortium, Philadelphia* 4.1 (2003), p. 34.

[20] Sergey Edunov, Alexei Baevski, and Michael Auli. "Pre-trained language model representations for language generation". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 4052–4059. DOI: 10.18653/v1/N19-1409. URL: https://aclanthology.org/N19-1409.

[21] Matthew E. Peters et al. "Deep contextualized word representations". In: *CoRR* abs/1802.05365 (2018). arXiv: 1802.05365. URL: http://arxiv.org/abs/1802.05365.

[22] Ondřej Bojar et al. "Findings of the 2018 Conference on Machine Translation (WMT18)". In: *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*. Belgium, Brussels: Association for Computational Linguistics, Oct. 2018, pp. 272–303. DOI: 10.18653/v1/W18-6401. URL: https://aclanthology.org/W18-6401.

[23] Abhay Shukla et al. *Legal Case Document Summarization: Extractive and Abstractive Methods and their Evaluation*. 2022. arXiv: 2210.07544 [cs.CL].

[24] Iz Beltagy, Matthew E. Peters, and Arman Cohan. "Longformer: The Long-Document Transformer". In: *CoRR* abs/2004.05150 (2020). arXiv: 2004.05150. URL: https://arxiv.org/abs/2004.05150.

[25] Long Ouyang et al. "Training language models to follow instructions with human feedback". In: *Advances in Neural Information Processing Systems*. Ed. by S. Koyejo et al. Vol. 35. Curran Associates, Inc., 2022, pp. 27730–27744. URL: https://proceedings.neurips.cc/paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf.

[26] Aashish Ghimire, Raj Shrestha, and John Edwards. "Too Legal; Didn't Read (TLDR): Summarization of Court Opinions". In: *2023 Intermountain Engineer-*

ing, Technology and Computing (IETC). 2023, pp. 164–169. DOI: 10.1109/IETC57902.2023.10152119.

[27] Yue Huang et al. "A High-Precision Two-Stage Legal Judgment Summarization". In: *Mathematics* 11.6 (2023). ISSN: 2227-7390. DOI: 10.3390/math11061320. URL: https://www.mdpi.com/2227-7390/11/6/1320.

[28] Hangbo Bao et al. "UniLMv2: Pseudo-Masked Language Models for Unified Language Model Pre-Training". In: *Preprint*. 2020.

[29] Filippo Galgani. *Legal Case Reports*. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5ZS41. 2012.

[30] Yang Liu and Mirella Lapata. "Text Summarization with Pretrained Encoders". In: *CoRR* abs/1908.08345 (2019). arXiv: 1908.08345. URL: http://arxiv.org/abs/1908.08345.

[31] Zhenrong Deng et al. "A Two-stage Chinese Text Summarization Algorithm Using Keyword Information and Adversarial Learning". In: *Neurocomputing* 425 (Mar. 2020). DOI: 10.1016/j.neucom.2020.02.102.

[32] Ramesh Nallapati, Feifei Zhai, and Bowen Zhou. "SummaRuNNer: A Recurrent Neural Network based Sequence Model for Extractive Summarization of Documents". In: *CoRR* abs/1611.04230 (2016). arXiv: 1611.04230. URL: http://arxiv.org/abs/1611.04230.

[33] Priyanka Prabhakar, Deepa Gupta, and Peeta Basa Pati. "Abstractive Summarization of Indian Legal Judgments". In: *2022 OITS International Conference on Information Technology (OCIT)*. 2022, pp. 256–261. DOI: 10.1109/OCIT56763.2022.00056.

[34] Colin Raffel et al. "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer". In: *CoRR* abs/1910.10683 (2019). arXiv: 1910.10683. URL: http://arxiv.org/abs/1910.10683.

[35] Yuxin Huang et al. "Legal public opinion news abstractive summarization by incorporating topic information". In: *International Journal of Machine Learning and Cybernetics* 11.9 (Sept. 2020), pp. 2039–2050. ISSN: 1868-808X. DOI: 10.1007/s13042-020-01093-8. URL: https://doi.org/10.1007/s13042-020-01093-8.

[36] Ingo Glaser, Sebastian Moser, and Florian Matthes. "Summarization of German Court Rulings". In: *Proceedings of the Natural Legal Language Processing Workshop 2021*. Punta Cana, Dominican Republic: Association for Computational

Linguistics, Nov. 2021, pp. 180–189. DOI: 10.18653/v1/2021.nllp-1.19. URL: https://aclanthology.org/2021.nllp-1.19.

[37] Kyunghyun Cho et al. "Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation". In: *CoRR* abs/1406.1078 (2014). arXiv: 1406.1078. URL: http://arxiv.org/abs/1406.1078.

[38] Jiyoung Yoon et al. "Abstractive Summarization of Korean Legal Cases using Pretrained Language Models". In: *2022 16th International Conference on Ubiquitous Information Management and Communication (IMCOM)*. 2022, pp. 1–7. DOI: 10.1109/IMCOM53663.2022.9721808.

[39] Daniel Núñez-Robinson, Jose Talavera-Montalto, and Willy Ugarte. "A Comparative Analysis on the Summarization of Legal Texts Using Transformer Models". In: *Advanced Research in Technologies, Information, Innovation and Sustainability*. Ed. by Teresa Guarda, Filipe Portela, and Maria Fernanda Augusto. Cham: Springer Nature Switzerland, 2022, pp. 372–386. ISBN: 978-3-031-20319-0.

[40] Anastassia Kornilova and Vladimir Eidelman. "BillSum: A Corpus for Automatic Summarization of US Legislation". In: *Proceedings of the 2nd Workshop on New Frontiers in Summarization*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 48–56. DOI: 10.18653/v1/D19-5406. URL: https://aclanthology.org/D19-5406.

[41] Luyang Huang et al. "Efficient Attentions for Long Document Summarization". In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online: Association for Computational Linguistics, June 2021, pp. 1419–1436. DOI: 10.18653/v1/2021.naacl-main.112. URL: https://aclanthology.org/2021.naacl-main.112.

[42] Sheetal Takale et al. "Legal Data Assistive Tool Using Deep-Learning". In: *2023 Third International Conference on Secure Cyber Computing and Communication (ICSCCC)*. 2023, pp. 791–796. DOI: 10.1109/ICSCCC58608.2023.10176784.

[43] Thiago Dal Pont et al. *Legal Summarisation through LLMs: The PRODIGIT Project*. 2023. arXiv: 2308.04416 [cs.CL].

[44] Gabriele Sarti and Malvina Nissim. *IT5: Large-scale Text-to-text Pretraining for Italian Language Understanding and Generation*. 2022. arXiv: 2203.03759 [cs.CL].

[45] OpenAI. *GPT-4 Technical Report*. 2023. arXiv: 2303.08774 [cs.CL].

[46] Diego de Vargas Feijo and Viviane P. Moreira. "Improving abstractive summarization of legal rulings through textual entailment". In: *Artificial Intelligence and Law* 31.1 (Mar. 2023), pp. 91–113. ISSN: 1572-8382. DOI: 10.1007/s10506-021-09305-4. URL: https://doi.org/10.1007/s10506-021-09305-4.

[47] Diego de Vargas Feijó and Viviane Pereira Moreira. "RulingBR: A Summarization Dataset for Legal Texts". In: *Computational Processing of the Portuguese Language*. Ed. by Aline Villavicencio et al. Cham: Springer International Publishing, 2018, pp. 255–264. ISBN: 978-3-319-99722-3.

[48] Batuhan Baykara and Tunga Güngör. "Turkish abstractive text summarization using pretrained sequence-to-sequence models". In: *Natural Language Engineering* (2022), pp. 1–30. DOI: 10.1017/S1351324922000195.

[49] Batuhan Baykara and Tunga Güngör. "Abstractive Text Summarization and New Large-Scale Datasets for Agglutinative Languages Turkish and Hungarian". In: *Lang. Resour. Eval.* 56.3 (Sept. 2022), pp. 973–1007. ISSN: 1574-020X. DOI: 10.1007/s10579-021-09568-y. URL: https://doi.org/10.1007/s10579-021-09568-y.

[50] Stefan Schweter. *BERTurk - BERT models for Turkish*. Version 1.0.0. Apr. 2020. DOI: 10.5281/zenodo.3770924. URL: https://doi.org/10.5281/zenodo.3770924.

[51] Yinhan Liu et al. "Multilingual Denoising Pre-training for Neural Machine Translation". In: *CoRR* abs/2001.08210 (2020). arXiv: 2001.08210. URL: https://arxiv.org/abs/2001.08210.

[52] Telmo Pires, Eva Schlinger, and Dan Garrette. "How Multilingual is Multilingual BERT?" In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, July 2019, pp. 4996–5001. DOI: 10.18653/v1/P19-1493. URL: https://aclanthology.org/P19-1493.

[53] Linting Xue et al. "mT5: A massively multilingual pre-trained text-to-text transformer". In: *CoRR* abs/2010.11934 (2020). arXiv: 2010.11934. URL: https://arxiv.org/abs/2010.11934.

[54] Thomas Scialom et al. "MLSUM: The Multilingual Summarization Corpus". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, Nov. 2020, pp. 8051–8067. DOI: 10.18653/v1/2020.emnlp-main.647. URL: https://aclanthology.org/2020.emnlp-main.647.

[55] Erol Gödur. "Hukuk Metinleri için Anahtar Kelime Kullanımı ile Otomatik Özetleme". In: *Rahva Teknik ve Sosyal Araştırmalar Dergisi* 1.1 (2021), pp. 24–36.

[56] Güneş Erkan and Dragomir R. Radev. "LexPageRank: Prestige in Multi-Document Text Summarization". In: *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*. Barcelona, Spain: Association for Computational Linguistics, July 2004, pp. 365–371. URL: https://aclanthology.org/W04-3247.

[57] Jacob Devlin et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *CoRR* abs/1810.04805 (2018). arXiv: 1810.04805. URL: http://arxiv.org/abs/1810.04805.

[58] Tiancheng Tang, Xinhuai Tang, and Tianyi Yuan. "Fine-Tuning BERT for Multi-Label Sentiment Analysis in Unbalanced Code-Switching Text". In: *IEEE Access* 8 (2020), pp. 193248–193256. DOI: 10.1109/ACCESS.2020.3030468.

[59] Görkem Göknar. *Turkish GPT2 Model Finetuned*. https://huggingface.co/gorkemgoknar/gpt2-small-turkish. Accessed: 2023-05-17. 2020.

[60] Mike Lewis et al. "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension". In: *CoRR* abs/1910.13461 (2019). arXiv: 1910.13461. URL: http://arxiv.org/abs/1910.13461.

[61] Abdullatif Köksal. *Pre-trained Word2Vec Model for Turkish*. 2020. URL: https://github.com/akoksal/Turkish-Word2Vec.

[62] Derek Miller. "Leveraging BERT for Extractive Text Summarization on Lectures". In: *CoRR* abs/1906.04165 (2019). arXiv: 1906.04165. URL: http://arxiv.org/abs/1906.04165.