Ministry of Science and Higher Education of the Republic of Kazakhstan

Suleyman Demirel University

Kaisar Barlybay

Question Answering system on Regulatory Documents

THESIS

Presented in Partial Fulfilment for the

Master of Technical Sciences Degree in Computer Science

(degree code: 7M06102)

Department of Computer Science

Faculty of Engineering and Natural Sciences

Supervisor: **Dauren Ayazbayev**

Kaskelen 2023

Suleyman Demirel University
Faculty of Engineering and Natural Sciences
Department of Computer Science

Dean of Faculty

Associate Professor

PhD Azamat Zhamanov

«06» 06 2023

Topic of the thesis:

Question Answering system on Regulatory Documents

Thesis submitted as part of the requirements for the award of the MSc in "7M06102 - Computer Science" SDU, 2021-2023

Head of Department _____ Assistant Professor, PhD Mukash Zh.

Academic Supervisor _____ Dauren Ayazbayev

Master student _____ Kaisar Barlybay

Kaskelen 2023

Ministry of Science and Higher Education of the Republic of Kazakhstan

Suleyman Demirel University

Kaisar Barlybay

Question Answering system on Regulatory Documents

THESIS

Presented in Partial Fulfilment for the

Master of Technical Sciences Degree in Computer Science

(degree code: 7M06102)

Department of Computer Science

Faculty of Engineering and Natural Sciences

Supervisor: **Dauren Ayazbayev**

Kaskelen 2023

Suleyman Demirel University
Faculty of Engineering and Natural Sciences
Department of Computer Science

Dean of Faculty

Associate Professor

PhD Azamat Zhamanov

_____

« _____ »_____ 2023

Topic of the thesis:

Question Answering system on Regulatory Documents

Thesis submitted as part of the requirements for the award of the MSc in "7M06102 - Computer Science" SDU, 2021-2023

Head of Department _____ Assistant Professor, PhD Mukash Zh.

Academic Supervisor _____ Dauren Ayazbayev

Master student _____ Kaisar Barlybay

Kaskelen 2023

Declaration

I confirm that this is my own work and the use of all material from other sources has been properly and fully acknowledged.

Kaisar Barlybay

2023

Abstract

The domain of legal text processing in the Kazakh language is currently underserved, presenting a unique challenge due to its specialized language and the relative scarcity of computational resources dedicated to it. This thesis explicitly identifies the problem: the need for an efficient model to process, understand, and generate meaningful insights from Kazakh legal texts.

Addressing this problem, the thesis proposes a solution by developing and evaluating bespoke language models pre-trained on a vast corpus of Kazakh legal documents. The study begins with the assembly of a corpus, which comprises over 315 million words from Kazakh legal texts, alongside a benchmark dataset of 2500 multiple-choice questions for civil service examinations in Kazakhstan.

Three language models based on the BERT architecture are then pre-trained. Among these, one model is pre-trained entirely from scratch. To emulate a real-world application in the legal domain, the performance of these models is assessed using the multiple-choice question-answering task.

The BERT base model pre-trained from scratch, leveraging both Masked Language Modeling (MLM) and Next Sentence Prediction (NSP) tasks, achieves an accuracy of 56.11%. This result underlines the potential of custom pre-training strategies on domain-specific corpora for enhancing the performance of language models in specialized areas.

In conclusion, this research represents a significant advancement in using AI for legal text processing in the Kazakh language. It presents a promising solution to the problem, paving the way for more efficient and informed decision-making processes in legal and civil service settings.

Аңдатпа

Қазіргі уақытта қазақ тіліндегі заңдық мәтінді өңдеу саласы жеткіліксіз қызмет етуде, бұл оның мамандандырылған тіліне және оған арналған есептеу ресурстарының салыстырмалы тапшылығына байланысты ерекше қиындық тудырады. Бұл дипломдық жұмыс проблеманы нақты анықтайды: қазақ заң мәтіндерін өңдеу, түсіну және мағыналы түсініктерді қалыптастыру үшін тиімді үлгінің қажеттілігі.

Бұл мәселені қарастыра отырып, дипломдық жұмыс қазақстандық құқықтық құжаттардың кең корпусында алдын ала дайындалған тапсырыстық тіл үлгілерін әзірлеу және бағалау арқылы шешуді ұсынады. Зерттеу Қазақстандағы мемлекеттік қызмет емтихандарына арналған 2500 көп таңдаулы сұрақтардан тұратын эталондық деректер жиынтығымен қатар қазақ заң мәтіндерінен 315 миллионнан астам сөзден тұратын корпусты құрастырудан басталады.

Содан кейін BERT архитектурасына негізделген үш тіл үлгісі алдын ала дайындалады. Олардың ішінде бір модель толығымен нөлден алдын ала дайындалған. Құқықтық домендегі нақты әлемдегі қолданбаны эмуляциялау үшін осы үлгілердің өнімділігі бірнеше таңдаулы сұраққа жауап беру тапсырмасы арқылы бағаланады.

Маскаланған тілді модельдеу (MLM) және келесі сөйлемді болжау (NSP) тапсырмаларын қолдана отырып, нөлден бастап алдын ала дайындалған BERT базалық үлгісі 56,11% дәлдікке жетеді. Бұл нәтиже мамандандырылған аймақтардағы тіл үлгілерінің өнімділігін арттыру үшін доменге тән корпуста теңшелетін алдын ала оқыту стратегияларының әлеуетін көрсетеді.

Қорытындылай келе, бұл зерттеу қазақ тіліндегі заңды мәтінді өңдеу үшін AI қолданудағы айтарлықтай ілгерілеушілікті білдіреді. Ол мәселенің перспективалық шешімін ұсынып, құқықтық және мемлекеттік қызмет орындарында тиімдірек және негізделген шешім қабылдау процестеріне жол ашады.

Аннотация

Область обработки юридических текстов на казахском языке в настоящее время недостаточно изучена, что представляет собой уникальную проблему из-за его специализированного языка и относительной нехватки вычислительных ресурсов, предназначенных для этого. В этом тезисе четко обозначена проблема: потребность в эффективной модели для обработки, понимания и получения осмысленных выводов из казахстанских юридических текстов.

Для решения этой проблемы в диссертации предлагается решение путем разработки и оценки индивидуальных языковых моделей, предварительно обученных на обширном корпусе казахских юридических документов. Исследование начинается со сборки корпуса, который включает более 315 миллионов слов из казахстанских юридических текстов, наряду с эталонным набором данных из 2500 вопросов с несколькими вариантами ответов для экзаменов на государственную службу в Казахстане.

Затем предварительно обучаются три языковые модели, основанные на архитектуре BERT. Среди них одна модель предварительно обучена полностью с нуля. Чтобы эмулировать реальное приложение в юридической области, производительность этих моделей оценивается с помощью задачи с множественным выбором ответов на вопросы.

Базовая модель BERT, предварительно обученная с нуля с использованием задач моделирования маскированного языка (MLM) и прогнозирования следующего предложения (NSP), обеспечивает точность 56,11%. Этот результат подчеркивает потенциал пользовательских стратегий предварительного обучения на предметно-ориентированных корпусах для повышения производи-

тельности языковых моделей в специализированных областях.

В заключение, это исследование представляет собой значительный прогресс в использовании ИИ для обработки юридических текстов на казахском языке. Он представляет собой многообещающее решение проблемы, прокладывая путь к более эффективным и обоснованным процессам принятия решений в юридических и государственных учреждениях.

Abbreviations

- BERT - Bidirectional Encoder Representations from Transformers
- SCOTUS - Supreme Court of the United States
- ECtHR - European Court of Human Rights
- NLP - Natural Language Processing
- ADTs - AdaBoosted decision trees
- NLG - Natural Language Generation
- CRF - Conditional Random Field
- CUAD - Contract Understanding Atticus Dataset
- BiLSTM - Bidirectional Long Short-Term Memory
- NER - Named Entity Recognition
- CRFs - Conditional Random Fields

Table of Contents

Declaration **i**

Abstract **ii**

Аңдатпа iii

Аннотация v

List of Abbreviations **vii**

1 Introduction **1**

1.1 Motivation 1

1.2 Aims and Objectives 2

1.3 Thesis Outline 3

2 Related work **6**

2.1 Legal Judgment Prediction and Decision Making 6

2.2 Legal Information Extraction and Text Analysis 8

2.3 Legal Question Answering Systems 10

2.4 Application of Pretrained Models in Legal Domain 12

2.5 Legal System Digitalization 14

3 Implementation of the domain language model **18**

3.1 Methodology 20

3.1.1 Masked Language Model (MLM) 20

3.1.2 Next Sentence Prediction (NSP) 20

3.2 Pre-training . . . 21
3.3 Pre-training corpora . . . 22
3.3.1 Scraping of the Kazakh Legal Documents Corpus . . . 23
3.3.2 Text Length Distribution and Frequent Lexical Patterns in Kazakh Legal Corpus . . . 24
3.3.3 Preparing the corpus of Kazakh legal documents for pre-training . . . 26

4 Evaluation **30**
4.1 Multiple Choice Question Answering Dataset . . . 31
4.1.1 Lexical Analysis of the Kazakh MCQA Dataset . . . 32
4.2 Model configuration . . . 35
4.3 Dataset encoding for MCQA . . . 36
4.4 Downstream task results . . . 38
4.5 Discussion . . . 40

5 Implications and Future Work **43**
5.1 Implications . . . 43
5.2 Future Work . . . 44

6 Conclusion **46**

Bibliography **48**

Chapter 1

Introduction

1.1 Motivation

With the advent of the digital age and advancements in technology, the field of law has witnessed an increasing shift towards digitization of legal services. Legal research and interpretation, often seen as complex and time-consuming, is one such area where technology can bring significant advantages. The ability to quickly and accurately find answers to legal questions can transform legal practices, making them more efficient and accessible [1].

While several question-answering systems exist for legal acts in English and other widely spoken languages, there is a significant gap when it comes to the Kazakh language. The Kazakh language, spoken by millions in Kazakhstan and other Central Asian countries, has a rich legal heritage with an array of regulatory legal acts. However, the complex legal terminology, compounded by the linguistic nuances of the Kazakh language, often presents a barrier to comprehension and application of these acts [2].

The lack of a specialized tool to assist with understanding and interpreting these legal acts is a notable gap in the legal tech landscape. Such a system can be invaluable for legal practitioners, students, and the general public in understanding legal provisions and their implications [3].

Furthermore, as Kazakhstan continues to modernize its legal system and

seeks to provide increased transparency and accessibility to its citizens, the need for such a system becomes even more critical. A question-answering system for regulatory legal acts in the Kazakh language would provide an essential tool in this drive towards legal digitization and transparency [4].

Lastly, the development of such a system could also provide valuable insights for similar work in other languages, particularly those languages that have not yet seen extensive research and development in the legal tech field. The development of a Kazakh language system could therefore be a stepping stone for advancements in legal tech for other under-researched languages [5], [6].

1.2 Aims and Objectives

The aim of this study is primarily focused on the creation of a language model in the Kazakh language specifically for the legal domain. This model can be used to form the core of a question-answering system in the legal domain. Here, we outline the main tasks of this work and the related specific objectives:

1. **Collection of Kazakh Text Corpus for Pre-training**: The foundation for training an efficient language model lies in the collection of a significant corpus of text data. Therefore, our first objective is to gather an extensive corpus of Kazakh legal documents, which will serve as the base for the pre-training of our BERT model. This comprehensive collection is expected to encompass a broad range of legal contexts and terminologies, enabling the subsequent model to understand and generate legal language effectively.

2. **Collection of Downstream Task Dataset**: Parallel to the collection of text corpus, our next objective is to compile a robust dataset for the downstream task - Multiple Choice Question Answering (MCQA). This dataset will play a pivotal role in fine-tuning the pre-trained language model and evaluating its performance in handling real-world tasks specific to the legal domain.

3. **Exploration of Existing Language Models**: Prior to pre-training

our BERT model, we aim to investigate the effectiveness of currently available language models on the selected downstream task. This will provide valuable insights into their capabilities and shortcomings, offering a benchmark for comparison and a pathway for improving our model's performance.

4. **Pre-training of BERT Model**: Armed with a substantial text corpus and an understanding of existing models, our next objective is to pre-train the BERT model using Next Sentence Prediction (NSP) and Masked Language Model (MLM) tasks. This step will enable the model to understand the contextual relationships in the Kazakh legal language and its intricacies.

5. **Evaluation of Pre-trained BERT Model**: After the pre-training phase, we will assess the model's performance on the downstream task. This will not only provide us with an understanding of the model's applicability and efficiency but also highlight areas where improvements may be needed.

6. **Comparison of Results**: Our final objective is to compare the performance of our pre-trained BERT model with that of existing models. This will allow us to measure our contributions and advancements in this field. The outcome of this comparison will validate our methodology and provide insights for future research directions.

The completion of these objectives will culminate in the creation of a full-fledged language model in the Kazakh language for the legal domain, providing a significant contribution to the ongoing development of automated legal assistance systems.

1.3 Thesis Outline

The structure of the thesis is as follows:

The thesis begins with the Introduction chapter, which sets the stage for the whole work. This section provides the Motivation behind the study, explaining

the problem it seeks to address and the significance of finding a solution. This chapter also delineates the Aims and Objectives of the research, clearly stating what the thesis hopes to achieve. The Thesis Outline section of this chapter offers a summary of the content and structure of the dissertation, giving the reader an idea of the journey ahead.

Following the introduction, the Preliminaries chapter delivers necessary background information to help readers understand the subsequent chapters. This includes crucial concepts, theories, and previous work that are related to the field of study.

The third chapter, Implementation of the Domain Language Model, dives into the core methodology used in the research. Here, the approach towards implementing the Masked Language Model (MLM) and Next Sentence Prediction (NSP) tasks are elaborated. The Pre-training section in this chapter covers the necessary steps that were taken to train the models using the collected datasets. It also discusses how the Pre-training Corpora were assembled, specifically detailing the Collection of the Kazakh Legal Documents Corpus and the process of Preparing the Corpus of Kazakh Legal Documents for Pretraining.

The fourth chapter, Evaluation, focuses on the application and assessment of the language models that were developed. This includes a comprehensive description of the Multiple Choice Question Answering Dataset and the specific Model Configuration used in the study. Additionally, it explains the Dataset Encoding for MCQA, detailing how the dataset was prepared to be used in the models. This chapter concludes with a section on Downstream Task Results that presents and discusses the outcome of applying the developed models on the MCQA task. A Discussion section follows, providing a comprehensive interpretation and analysis of the results.

The fifth chapter, Implications and Future Work provides an overview of the implications that this research could have in the field and outlines potential directions for future research in this domain. This includes a subsection on Implications, detailing how the research findings could be applied in real-world situations, and a section on Future Work, where potential improvements,

modifications, and extensions of the research are suggested.

The final chapter, Conclusion, offers a summary of the entire study. It synthesizes the research questions, methodology, findings, implications, and suggestions for future work into a cohesive closing argument. It underscores the importance and contribution of the research and offers a reflective assessment of the work done.

In summary, this dissertation not only presents the development and evaluation of a specialized language model for the Kazakh language in the legal domain but also provides insights and directions for future work in this field.

Chapter 2

Related work

In the context of the evolving digital landscape, the application of artificial intelligence (AI) in various sectors, including the legal domain, has seen a significant increase. Particularly, natural language processing (NLP) techniques and machine learning algorithms have been instrumental in developing question-answering systems for legal act interpretation. This section provides a review of the relevant literature that forms the backbone of this research (Table 2.1).

2.1 Legal Judgment Prediction and Decision Making

Legal judgment prediction is a growing area of interest in legal informatics, with research broadly focusing on three main strands. These are human rights violation cases at the European Court of Human Rights (ECtHR), Chinese criminal cases, and cases at the Supreme Court of the United States (SCOTUS). Significant strides have been made utilizing natural language processing (NLP) and machine learning to aid in predicting legal judgments in these contexts (Table 2.1).

In terms of ECtHR cases, Aletras et al. [1] marked a pioneering effort in predicting outcomes based solely on the textual content of case files. The study formulated a binary classification task, demonstrating that the formal facts

of a case, captured in the textual content, are most significant in predicting judgments. This finding emphasized the potential of NLP in judicial decision-making and provided a foundation for subsequent studies.

Building upon Aletras et al.'s work, Chalkidis et al. [7] introduced a new English legal judgment prediction dataset from the ECtHR and evaluated a range of neural models. In this study, the potential biases of these models were explored, notably towards demographic information. To manage these biases, they employed data anonymization techniques and introduced a hierarchical version of BERT, a powerful NLP model, highlighting the relevance of the hierarchical structure inherent in legal documents.

The field of legal text processing has witnessed considerable growth, as evidenced by numerous studies and comprehensive reviews on the topic [8], [9], [10].

Medvedeva et al. [11] further leveraged NLP tools to predict judicial decisions. Despite a high average accuracy of 75% in predicting violations of nine articles of the European Convention on Human Rights, their model's performance decreased when applied to future cases based on past ones, averaging between 58% to 68% accuracy. Interestingly, they discovered that predictions could still achieve relatively high accuracy (65%) when made solely based on the surnames of the judges presiding over the case. This indicates that certain extrajudicial factors may also play significant roles in legal decisions.

In the realm of Chinese criminal cases, significant advancements have been made. Zhong et al. [12] introduced a topological multi-task learning framework, TOPJUDGE, formalizing the dependencies among various subtasks in legal judgment prediction. This approach underlined the interconnected nature of legal cases, where different aspects of a case are intrinsically linked. Yang et al. [13] further built on this multi-task learning concept by integrating word collocation features of fact descriptions into a neural network using an attention mechanism. This approach effectively improved the prediction accuracy in cases with similar fact descriptions but different penalties, demonstrating the complexity and nuance inherent in legal judgment prediction.

Shifting focus to the SCOTUS, research has been heavily influenced by the work of Martin-Katz et al. [14]. The study employed a time-evolving random forest classifier model to predict SCOTUS behavior in a generalized, out-of-sample context. The model achieved impressive accuracy rates of 70.2% at the case outcome level and 71.9% at the justice vote level, spanning over two centuries of cases (1816-2015).

The exploration of advanced machine learning methods to further improve SCOTUS prediction was championed by Kaufman [15]. The study utilized AdaBoosted decision trees (ADTs), outperforming existing predictive models. This research also applied the ADT approach to other prediction tasks, such as the onset of civil wars and county-level vote shares in U.S. presidential elections, demonstrating the versatility and potential of such machine learning techniques in the domain of prediction tasks.

Overall, these studies reflect the tremendous strides made in the realm of legal judgment prediction. From predicting human rights violations to criminal charges and SCOTUS rulings, the research signifies the potential for NLP and machine learning to aid and transform traditional legal procedures. However, it also underscores the intricate nature of the field, where simple textual analysis may not always suffice and other case-related factors and data types must be considered.

2.2 Legal Information Extraction and Text Analysis

Legal Information Extraction and Text Analysis is another significant field within legal informatics that not only predicts court decisions but also interprets specific judgments and processes legal texts. Multiple studies have approached this field from various perspectives, underscoring the richness and complexity of legal language and documentation (Table 2.1).

In the realm of judgment interpretation, Ye et al. [16] introduced a novel application of natural language generation (NLG) in a task they termed COURT

VIEW GENeration. Their objective was to increase the interpretability of charge prediction systems and facilitate the automatic generation of legal documents. They accomplished this by creating a label-conditioned Seq2Seq model with attention, which generates court views conditioned on encoded charge labels. This work emphasized the necessity of not only predicting the legal outcome but also understanding the reasoning that leads to the decisions.

Building on the notion of explanation extraction, Chalkidis et al. [17] explored rationale extraction at a paragraph level in multi-paragraph structured court cases. They released a new dataset of European Court of Human Rights cases with paragraph-level rationale annotations. The study also introduced a new constraint, singularity, which improved the quality of rationales, indicating the importance of structure and constraint imposition in extracting meaningful information from legal texts.

Branting [18] further extended the exploration of explanation extraction by presenting two methods for outcome prediction in legal decision-support systems. One of these methods, semi-supervised case annotation, identifies explanatory textual patterns related to case decisions based on structural and semantic regularities in case corpora, reinforcing the importance of incorporating structural features in the analysis of legal texts.

In a related line of research, researchers have used NLP and AI techniques to study legal text processing. For example, Zhong et al. [12] introduced a topological multi-task learning framework, TOPJUDGE, which formalizes the dependencies among subtasks in legal judgment prediction, demonstrating the interconnected nature of various aspects in legal cases. Chalkidis et al. [7] expanded on this, creating an English dataset of legal judgment predictions and employing various neural models for processing.

Valvoda [19] introduced an information-theoretic approach to understand the influence of precedents in legal decisions. The research suggested that the arguments of precedents could be more influential than the precedent's facts in deciding the case's outcome, illustrating the importance of understanding and incorporating legal reasoning in the study of legal texts.

With regard to the application of NLP in contract law, Chalkidis et al. [20] worked on detecting contractual obligations and prohibitions using hierarchical BILSTM. Their approach outperformed the flat model as it provided a broader discourse view, thus emphasizing the advantage of hierarchical models in capturing the structured nature of legal texts.

Similarly, Hendrycks et al. [21] introduced the Contract Understanding Atticus Dataset (CUAD), a dataset for legal contract review. The task involved highlighting key sections of a contract for human review, a task where Transformer models showed promising performance, albeit with scope for improvement.

In a slightly different direction, Leitner et al. [22] utilized Named Entity Recognition (NER) in German legal documents. They developed a dataset with 19 manually annotated semantic classes and applied Conditional Random Fields (CRFs) and bidirectional Long-Short Term Memory Networks (BiLSTMs) for NER tasks, which achieved impressive performance scores.

These studies showcase the increasing use and importance of NLP and AI in legal information extraction and text analysis. They reflect the growing focus on not just predicting but also understanding and explaining legal judgments and extracting valuable information from legal texts such as contracts and court cases.

2.3 Legal Question Answering Systems

Legal Question Answering (QA) Systems have significantly advanced over the past decades, moving from early rule-based systems to sophisticated machine learning-based models. These systems aim to provide accurate and efficient responses to legal queries, promoting a better understanding of legal rights among the public and facilitating research (Table 2.1).

QA systems gained recognition after Voorhees's discussion of the TREC-8 QA Track in 2000, which marked the first large-scale evaluation of domain-independent QA systems. The authors investigated whether the document

retrieval evaluation methodology was applicable to other natural language processing (NLP) tasks and found that creating a reusable QA test collection is more complex than document retrieval, as there are no equivalents to document identifiers in QA tasks (Voorhees2000). This pioneering work laid the groundwork for future developments in the QA field, including in the legal domain.

Subsequently, the field witnessed significant progress with the introduction of BERT (Bidirectional Encoder Representations from Transformers) by Devlin et al. [23]. Designed to pre-train deep bidirectional representations from unlabeled text, BERT revolutionized the language understanding capabilities of QA systems. Fine-tuning the pre-trained BERT model with an additional output layer created state-of-the-art models for various tasks, demonstrating its adaptability and efficiency.

The capabilities of QA systems were further advanced with the introduction of GPT by Alec et al. [24]. They showed that language models could begin to learn tasks without any explicit supervision when trained on a massive dataset like WebText. The paper revealed a promising path towards developing language processing systems that learn to perform tasks through naturally occurring demonstrations.

In the legal domain, Kim et al. [25] developed a QA system that combined legal information retrieval and textual entailment, with a focus on paraphrasing and sentence-level analysis of queries and legal statutes. The system was evaluated using the training data from the COLIEE-2016 competition, showcasing its potential to answer yes/no questions from Japanese legal bar exams effectively. This work represented a significant step towards the development of legal QA systems that can handle complex, statute-based queries.

Ravichander et al. [26] took this further by introducing PrivacyQA, a corpus consisting of questions about the privacy policies of mobile applications. The paper highlighted the challenges of question answerability and emphasized the need for improved system performance in privacy policy-related QA. This work underscored the necessity of specialized QA systems to handle the complex and often opaque privacy policies, a topic of growing importance in the digital age.

Kien et al. [27] focused on answering legal questions at the article level and proposed a model that used convolutional neural networks and attention mechanisms for text representation. This model performed well on an annotated corpus of Vietnamese legal questions, demonstrating its efficacy in retrieving relevant legal articles for given legal queries. Their work highlighted the potential of using advanced neural network models and attention mechanisms in QA systems for legal queries.

The study [28] focused on the evaluation of BERT models and their derivatives for the Multiple Choice Question Answering (MCQA) task in the legal domain for the Kazakh language. The study found that the distilbert-base-multilingual-cased model showed the highest accuracy at 46.27%, while the bert-base-uncased model showed the lowest performance. These results highlight the importance of multilingual learning for effective cross-language learning and underscore the potential of BERT models in tasks involving less resource-intensive languages, such as Kazakh.

QA systems, especially in the legal domain, have seen considerable advances over the years, fueled by developments in NLP, deep learning, and an increasing emphasis on public legal education. However, challenges remain, and the development of more sophisticated legal QA systems capable of handling the intricacies of legal language and concepts is an ongoing pursuit in the field.

2.4 Application of Pretrained Models in Legal Domain

The application of Transformer-based language models pre-trained on legal corpora has marked a significant advancement in legal Natural Language Processing (NLP). This advancement has mainly been triggered by several studies that have focused on fine-tuning these pre-trained models for legal-specific tasks such as information extraction, case prediction, and legal question-answering (Table 2.1).

A key development in this space is LEGAL-BERT, a specialized version of

BERT introduced by Chalkidis et al. [29] tailored for legal NLP tasks, computational law, and legal tech applications. The authors conducted a comprehensive examination of different strategies for applying BERT in specialized domains, yielding valuable insights into optimizing the model's performance. They proposed a more expansive hyper-parameter search space for fine-tuning, suggesting the potential for tuning these models according to specific task requirements and thereby optimizing their performance in specialized applications.

Adding to the discourse, Zheng et al. [30] presented the CaseHOLD dataset, composed of more than 53,000 multiple-choice questions aimed at identifying the relevant holding of a cited case. This work addressed a crucial question: when does pretraining help? Their findings underscored the importance of domain-specific pretraining with a custom legal vocabulary, which showed significant performance gains in legal NLP tasks. Interestingly, the level of performance improvement was found to be directly proportional to the domain specificity of the task. This study supports the idea that domain-specific pretraining can significantly enhance model performance, particularly for tasks with high domain specificity.

Extending this trend of domain-specific pretraining to non-English languages, Xiao et al. [31] introduced Lawformer, a model specifically designed for understanding Chinese legal long documents. This model, based on the Longformer architecture, demonstrated impressive improvements on various LegalAI tasks, including judgment prediction, similar case retrieval, legal reading comprehension, and legal question answering. This work expands the reach of domain-specific pretrained models to multilingual legal corpora, which is crucial for legal NLP tasks in non-English jurisdictions.

These advancements underline the potential of domain-specific pretraining in the legal field. Models like LEGAL-BERT, complemented by the insights derived from studies like CaseHOLD and Lawformer, are paving the way for more efficient and effective NLP applications in the legal domain. It's evident that the benefits of pretraining Transformer-based models on domain-specific corpora can be significant, particularly for specialized fields such as law, which are characterized by unique terminologies and complex linguistic structures.

2.5 Legal System Digitalization

The digitalization of legal systems and the use of natural language processing (NLP) in law are fields of ongoing research and application. In the context of the Kazakh language and legal system, though progress has been made in both areas, a comprehensive application combining these domains remains largely unexplored (Table 2.1).

The work of Yelibayeva et al. [32] represents a significant stride in Kazakh language processing. They introduced an ontological model specifically developed for identifying nominative word combinations in the Kazakh language. This approach aimed at facilitating semantic searches, Q&A systems, e-learning platforms, and other software applications intended for knowledge acquisition. While the potential application in the legal domain can be inferred, the research itself did not expound upon an explicit application in legal act interpretation. It does, however, lay essential groundwork for potential development in this direction.

Simultaneously, the progress in the digitalization of the Kazakh legal system has been the focus of several studies. Ondashuly et al. [33] explored this topic in depth, particularly in the context of the judicial system. They provided comprehensive definitions of "digitalization of the legal system" and "digitalization of the judicial system," and detailed the adoption of advanced IT technologies and electronic projects in contemporary legal proceedings. This study shows the technological advancements being integrated into the legal ecosystem of Kazakhstan, signalling the readiness and potential for incorporating more sophisticated tools, such as NLP-based applications, into the system.

However, while these studies have made significant strides in their respective domains, they have not yet addressed the intersection of these two areas. The potential for a dedicated question-answering system for legal acts in the Kazakh language, leveraging the advances in both Kazakh language processing and the digitalization of the country's legal system, remains unexplored. It presents an open field for researchers and developers, offering a chance to innovate and create applications that can significantly streamline legal processes and aid in

the understanding and interpretation of legal acts.

This thesis aims to fill this gap by developing a BERT based language model suitable for question-answering system tailored for regulatory legal acts in the Kazakh language. By bridging the intersection of Kazakh language processing and legal technology, this study intends to contribute significantly to the field.

Table 2.1: Classification and Contribution of Key Research Works

Paper	**Year**	**Key Contributions**
Legal Judgment Prediction and Decision Making		
Zhong et al. [12]	2018	Explores topological learning methods for legal judgement prediction
Chalkidis et al. [7]	2020	Discusses the application of neural networks for legal judgement prediction
Aletras et al. [1]	2016	Applies NLP techniques to predict ECHR decisions
Medvedeva et al. [11]	2020	Uses ML techniques to predict decisions of the ECHR
Yang et al. [13]	2019	Uses a multi-perspective bi-feedback network for predicting legal judgments
Kaufman et al [15].	2019	Improves Supreme Court forecasting using boosted decision trees
Martin Katz et al. [14]	2017	Proposes a general approach for predicting SCOTUS behavior
Branting [18]	2021	Discusses methods for scalable and explainable legal prediction
Legal Information Extraction and Text Analysis		
Ye et al. [16]	2018	Provides an interpretable model for charge prediction in criminal cases
Branting [18]	2021	Discusses methods for scalable and explainable legal prediction

Continued on next page

Table 2.1 – continued from previous page

Paper	Year	Key Contributions
Chalkidis et al. [17]	2021	Paragraph-level Rationale Extraction through Regularization: A case study on European Court of Human Rights Cases
Valvoda et al. [19]	2021	Performs an information-theoretic analysis of common law
Chalkidis et al. [8]	2019	Discusses the use of deep learning and word embeddings in law
Bommarito II et al. [10]	2021	Proposes LexNLP for processing and information extraction in legal texts
Zhong et al. [9]	2020	Provides a summary of LegalAI and NLP's benefits to the legal system
Chalkidis et al. [20]	2018	Applies hierarchical RNNs to extract obligations and prohibitions from text
Hendrycks et al. [21]	2021	Introduces CUAD, a dataset for legal contract review
Leitner et al. [22]	2019	Focuses on fine-grained NER in legal documents
Legal Question Answering Systems		
Voorhees [34]	2000	Discusses the construction of QA test collections
Devlin et al. [23]	2019	Proposes BERT, a pre-trained transformer model for language understanding
Alec et al. [24]	2019	Presents an extensive study of unsupervised multitask learners
Kim et al. [35]	2017	Uses paraphrasing and legal text analysis for answering bar exam questions
Ravichander et al. [26]	2019	Combines computational and legal perspectives for QA on privacy policies
Kien et al. [27]	2020	Develops a neural attentive model for legal QA

Continued on next page

Table 2.1 – continued from previous page

Paper	Year	Key Contributions
Chalkidis et al. [29]	2020	Proposes LEGAL-BERT, a variant of BERT for the legal domain
Barlybay et al. [28]	2020	Presents a comprehensive evaluation of various BERT models for the Multiple Choice Question Answering (MCQA) task in the legal domain for the Kazakh language
Application of Pretrained Models in Legal Domain		
Chalkidis et al. [36]	2020	Conducts an empirical study on large-scale multi-label text classification
Mencia et al. [37]	2007	Evaluates efficient multi-label classification algorithms for large-scale problems in the legal domain
Zheng et al. [38]	2021	Assesses the impact of pretraining and self-supervised learning in law
Xiao et al. [31]	2021	Proposes Lawformer, a pre-trained language model for Chinese legal long documents
Legal System Digitalization		
Ondashuly et al.[33]	2018	Discusses the digitalization of the legal system of Kazakhstan
Yelibayeva et al. [32]	2021	Presents an ontological model for extracting Kazakh language word combinations in NLP

Chapter 3

Implementation of the domain language model

The construction of a question-answering system tailored for regulatory legal acts in the Kazakh language involves a multi-faceted process that includes language understanding, information retrieval, and machine learning.

The implementation of the BERT language model for the Kazakh language in the legal domain involved several steps, grounded on the BERT architecture's specifications [23].

1. **Tokenizer vocabulary size - 30522**: This parameter indicates the size of the vocabulary that the BERT model was trained on. It refers to the number of the most frequent words in the language corpus that the tokenizer keeps. The tokenizer breaks down text into tokens or pieces that are understandable by the model. In this case, the vocabulary size of 30,522 means that the model is capable of understanding and generating 30,522 distinct tokens.

2. **Hidden size - 768**: The BERT model uses a transformer architecture, which employs a concept called "hidden layers" or "hidden states". These hidden layers capture information from the input data and pass it along to the next layer. The hidden size of 768 denotes the dimensionality of

the hidden states, or the size of the output vectors from each layer in the model. A higher number can capture more complex patterns but also requires more computational power and training time.

3. **Maximum sequence length - 512**: This parameter determines the maximum length of a sentence (in tokens) that the model can handle. Sentences longer than this will be truncated, while shorter sentences will be padded to match this length. For BERT, this is typically set to 512.

4. **Number of heads - 12**: This refers to the number of attention heads in the transformer model. The attention mechanism allows the model to focus on different parts of the input when producing an output. With 12 heads, the model can pay attention to 12 different parts of the input simultaneously, capturing various aspects of the context.

5. **Number of parameters - 110M**: This indicates the complexity of the model, with 110 million trainable parameters. These parameters are the parts of the model that are updated during training to minimize the difference between the model's predictions and the actual data.

The model was trained on a legal corpus in the Kazakh language, allowing it to understand and generate legal text accurately. The training process would involve feeding the model large amounts of legal text data, and adjusting the model's parameters to minimize the difference between its predictions and the actual data.

Once the model was trained, it could be used for various tasks in the legal domain, such as question answering, legal document summarization, or legal judgment prediction.

These models require significant computational resources for training, and the training process can be time-consuming. However, once trained, these models can provide highly accurate results in the legal domain.

3.1 Methodology

The BERT model is typically pre-trained using two types of unsupervised learning tasks: Masked Language Model (MLM) and Next Sentence Prediction (NSP).

3.1.1 Masked Language Model (MLM)

The MLM task randomly masks a percentage of the input tokens (15% in the original BERT paper) with the objective of predicting the original vocabulary id of the masked word based only on its context. Unlike traditional language modeling, MLM is bidirectional, which allows the model to capture context from both the left and the right of a masked token.

The masking process involves replacing the selected tokens with a special [MASK] token, a random token, or leaving the token unchanged. The final objective for the MLM task is to minimize the discrepancy between the predicted probability distribution of the masked tokens and the true distribution.

3.1.2 Next Sentence Prediction (NSP)

The NSP task is designed to help the model understand the relationship between two sentences, which is beneficial for tasks such as question answering and natural language inference. During pre-training, the model is presented with pairs of sentences (A and B) and must predict whether sentence B follows sentence A in the original document.

Approximately 50% of the pairs are actual consecutive sentences from the document, while the other 50% are random sentences from the corpus. The model is then trained to predict whether sentence B is a random sentence or a sentence that follows sentence A.

The combination of MLM and NSP during pre-training allows BERT to understand both the syntax and semantics of a language, as well as the relationship between sentences. Once pre-training is complete, the model can be fine-tuned

on a specific downstream task using supervised learning, such as sentiment analysis, named entity recognition, or question answering.

It's worth noting that newer transformer models like RoBERTa, a variant of BERT, have abandoned the NSP task during pre-training and instead focused on a more robust implementation of MLM, as they found the NSP task didn't contribute significantly to the model's performance on downstream tasks.

3.2 Pre-training

The models used and the corresponding pre-training procedures are as follows:

1. **bert-base-multilingual-uncased**: This model was taken as a basis and pre-trained using only the MLM task. The batch size for the training was set to 4 sequences, with each sequence containing 512 tokens, which amounts to 2048 tokens per batch. The model was trained for 68,000 steps, which corresponds to approximately 5 epochs over a 112 million word corpus.

2. **BERT original**: The original BERT model was pre-trained using both MLM and NSP tasks from scratch. The batch size for the training was also set to 4 sequences, with each sequence containing 512 tokens, which amounts to 2048 tokens per batch. The model was trained for 177,000 steps, which is approximately equivalent to 1.5 epochs over a 315 million word legal domain corpus.

3. **bert-base-multilingual-cased**: This model was taken as a basis and pre-trained using both MLM and NSP tasks. The batch size for the training was set to 4 sequences, with each sequence containing 512 tokens, which amounts to 2048 tokens per batch. The model was trained for 90,000 steps, which is approximately 0.4 epochs over a 315 million word legal domain corpus.

These models were pre-trained on a GPU with 10GB memory using the Pytorch library. The choice of the Pytorch library provides extensive modularity and control, which aids in better model tuning and optimization. Pre-training these

models on such a vast corpus of legal texts makes them potent tools for natural language understanding in the legal domain of the Kazakh language.

3.3 Pre-training corpora

The language model was pre-trained on two major corpora, specifically designed to encapsulate a comprehensive understanding of both general Kazakh language and specific legal terminology Table 3.1. This dual focus is aimed at ensuring the model's proficiency in understanding and generating accurate and relevant responses within the legal domain in the Kazakh language.

- **Mixed Corpus**: This corpus is a hybrid of two distinct corpora. The first part of this mixed corpus is the OSCAR corpus for the Kazakh language, which contains 110 million words. OSCAR (Open Super-large Crawled ALMAnaCH coRpus) is a massive multilingual corpus obtained by language classification and filtering of the Common Crawl corpus. The inclusion of this corpus ensures that the model is well-versed in general Kazakh language structures and can understand and generate everyday Kazakh text. The second part of this mixed corpus is a set of Kazakh civil codes, containing 2 million words. These civil codes ensure that the model gains exposure to formal legal language, terminologies, structures, and conventions used within the realm of civil law in Kazakhstan.
- **Kazakh Legal Documents Corpus**: This corpus consists of a substantial collection of Kazakh legal documents, amassing to approximately 315 million words. This vast corpus enables the model to delve deeper into the realm of legal language and terminology specific to the Kazakh legal system. It provides the model with a comprehensive understanding of a wide variety of legal documents, including court case transcriptions, legal opinions, statutes, regulations, contracts, and many more. The exposure to such a diverse range of legal documents aids the model in identifying patterns, relationships, and structures in legal texts, making it highly effective for various tasks in the legal domain.

Both these corpora contribute towards creating a powerful language model that

Pre-training corpora	
Corpus name	Number of words
OSCAR (Open Super-large Crawled Aggregated corpus) - Kazakh language	110×10^6
Kazakh civil codes	2×10^6
Kazakh regulatory documents	3.15×10^8

Table 3.1: Corpora used for BERT pre-training

can navigate and understand the complexities of the legal language, making it an effective tool for various legal NLP tasks in the Kazakh language.

3.3.1 Scraping of the Kazakh Legal Documents Corpus

The assembly of the Kazakh Legal Documents Corpus was a comprehensive task that involved using web scraping tools, specifically BeautifulSoup (bs4), to collect data from the open source of Kazakh legal documents, available at https://adilet.zan.kz.

The process started with downloading each legal document as a separate file. These files were designed such that each contained a single line that represented the contents of an entire document, ensuring that the documents could be efficiently processed later on.

After the individual files were created, they were amalgamated into a single .txt file. This was done by appending each file's content into the master .txt file, with a line break character used to denote the separation between different documents. The resulting file contains an impressive total of over 315 million words, showcasing the significant amount of legal information that was scraped.

The final step involved splitting the comprehensive corpus into a test corpus and a training corpus. The division was done in a ratio of 5 to 95, with the larger portion dedicated for training, following the common practice in machine learning. This strategy allows for robust training of the model on the majority of the data, while still retaining a significant portion for evaluation purposes, thereby ensuring that the model's performance can be effectively gauged.

The collection of the Kazakh Legal Documents Corpus involved a meticulous process of data scraping, file creation, and corpus division, resulting in a large-scale and invaluable resource for training language models in the legal domain of the Kazakh language.

3.3.2 Text Length Distribution and Frequent Lexical Patterns in Kazakh Legal Corpus

The analysis of the most frequently occurring unigrams, bigrams, and trigrams in the corpus of legal documents in the Kazakh language can provide interesting insights into the main themes and topics addressed in this corpus.

Firstly, the prominence of unigrams Figure 3.1 like 'state', 'Kazakhstan', 'provincial', 'republic', 'city', 'service', 'mayor', 'village', and 'education' suggests a strong focus on state-level matters and policies, especially those related to administrative divisions (such as provinces, cities, and villages), public services, and education.

Bigrams Figure 3.2 such as 'regional importance', 'republic of Kazakhstan', 'important city', 'district akim', 'rural district', 'public service', and 'service provision' reinforce the idea of the significant role of administrative divisions, governance, and public services in the corpus. An 'akim' is an administrative leader or a mayor in the Kazakh context, so the phrase 'district akim' indicates discussions around local leadership and governance. The term 'regional importance' suggests legal matters related to key regions or cities.

The analysis of trigrams Figure 3.3 provides further context. Phrases like 'city of regional importance', 'akim of rural district', 'city rural district', 'village village rural', and 'state policy in the field' hint at the consideration of urban and rural dynamics, as well as state policies pertaining to specific areas or fields. Furthermore, the appearance of trigrams like 'from the date of official publication' and 'to be put into effect from the beginning' indicates procedural aspects related to the implementation of policies or regulations.

This corpus seems to heavily concentrate on state and local governance, ad-

ministrative divisions, public services, and the procedures of policy implementation and enforcement. The appearance of 'education' as a key unigram also suggests that educational policies or regulations form an important part of the legal corpus in Kazakh. This analysis can be helpful in further research, modeling, and understanding the main focus of the legal documents in this corpus.

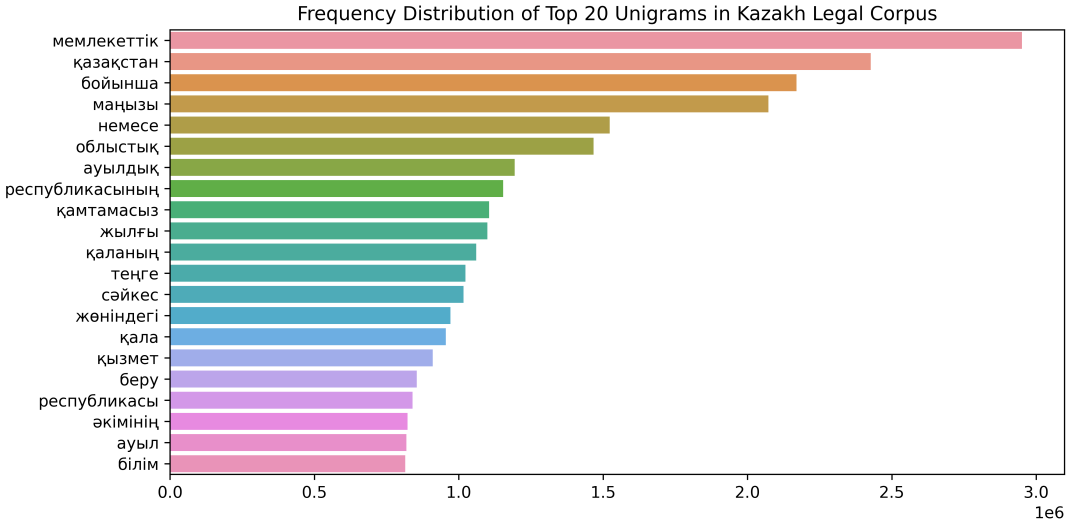

Figure 3.1: Frequency Distribution of Top 20 Unigrams in Kazakh Legal Corpus

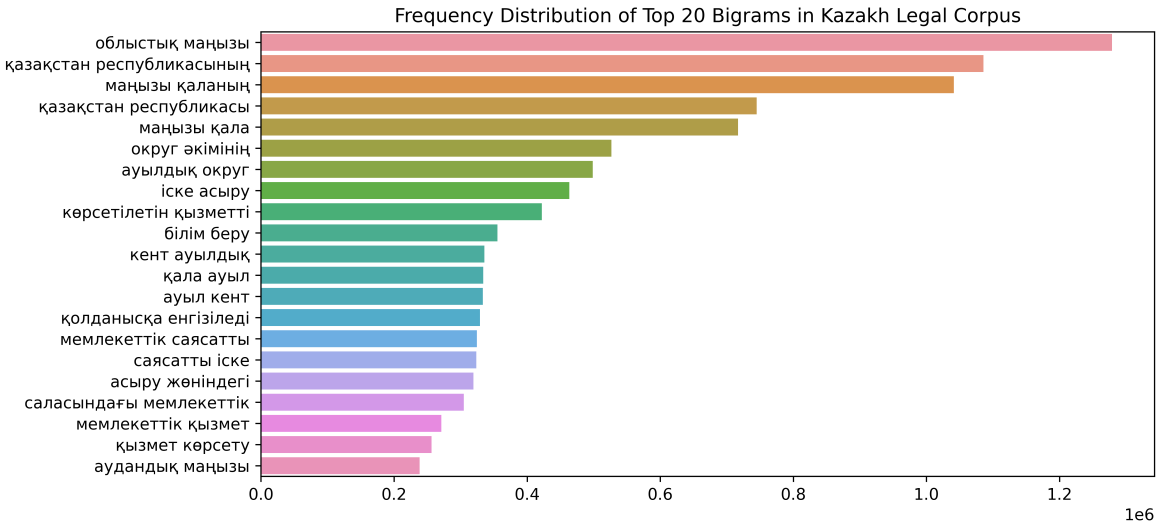

Figure 3.2: Frequency Distribution of Top 20 Bigrams in Kazakh Legal Corpus

The distribution of the number of documents to number of words in the text of the Kazakh legal corpus is skewed to the left (negative skew) Figure 3.4. The majority of the documents contain fewer words, with almost two-thirds (around 100,000 documents out of 152,966 documents) containing between 0 and 1000 words.

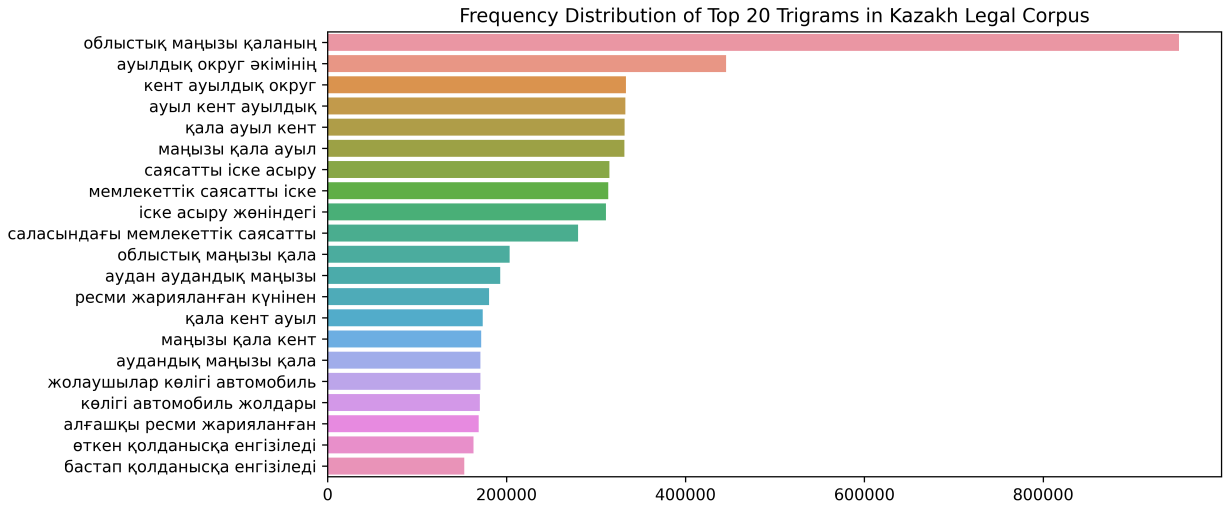

Figure 3.3: Frequency Distribution of Top 20 Trigrams in Kazakh Legal Corpus

A substantial number of documents, over 20,000, contain between 1000 to 2000 words, showing that a significant portion of the corpus consists of moderately lengthy documents. The number of documents that contain between 2000 and 3000 words is more than 10,000, which is less compared to the previous two categories but still considerable.

As the number of words increases, the number of documents that contain that many words decreases. Less than 2500 documents contain between 8000 and 10000 words, indicating that very lengthy documents are quite rare in the corpus.

This pattern of distribution is commonly observed in many natural language processing tasks where most documents are relatively short, and very long documents are less frequent.

3.3.3 Preparing the corpus of Kazakh legal documents for pre-training

Tokenizer training This phase of the project involved training a tokenizer using the BertWordPieceTokenizer. This tokenizer is particularly well-suited to the BERT architecture, as it employs WordPiece tokenization, a method that breaks down words into subwords and combines them into one token, if required.

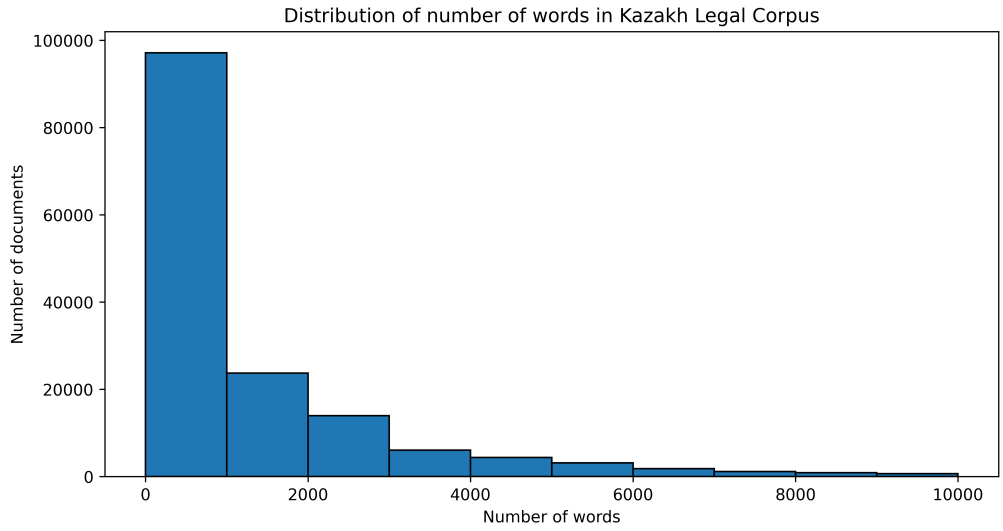

Figure 3.4: Distribution of number of words in Kazakh Legal Corpus

The training of the tokenizer was performed on the collected corpus of Kazakh legal documents, leveraging the extensive legal text data to learn the tokenization rules specific to the Kazakh language and the legal domain. This data-driven approach ensures that the tokenizer effectively captures the linguistic nuances present in the corpus, enhancing the accuracy of the tokenization process.

The maximum number of words (tokens) was set to 30,522 for the tokenizer. This choice balances the need for capturing sufficient lexical variety and complexity, while keeping the computational requirements manageable.

Moreover, to ensure the consistency of input length for downstream tasks and to manage computational demands, any document that exceeds 512 tokens in length would be truncated. While this may lead to some information loss for particularly long documents, it's a necessary compromise to keep the token sequences within a manageable length for processing by the BERT model.

Algorithm for obtaining features for NSP task The algorithm described is used for pre-training the BERT model on the Next Sentence Prediction (NSP) task. The goal of NSP is to allow the model to understand the relationship between two sentences, which is vital for many downstream tasks such as question answering and natural language inference. Here are the steps in the algorithm:

1. **Sentence Segmentation**: The documents from the Kazakh legal corpus are divided into individual sentences using the Natural Language Toolkit (NLTK), a popular library for linguistic data processing.

2. **Sentence Truncation**: Each sentence from each document undergoes truncation with a 10% chance. This is done to minimize any mismatch between the pre-training and fine-tuning stages, ensuring that the model is exposed to a variety of sentence lengths during training.

3. **Sentence Pair Formation**: To each sentence (now referred to as the 'first' sentence), a 'second' sentence is appended. This second sentence is separated from the first by a special [SEP] token. With a 50% probability, this second sentence is not the succeeding sentence from the original document, but a random sentence picked from elsewhere in the corpus. This forms a pair of sentences, half of which are consecutive sentences from the original document ('IsNext' label), and half are not ('NotNext' label).

4. **Tokenization and Padding/Truncation**: The final sequence (composed of the first sentence, the [SEP] token, and the second sentence) is then tokenized by the pre-trained tokenizer. If the tokenized sequence is shorter than 512 tokens, it is padded with the special [PAD] tokens to reach the required length. If it's longer, it is truncated to fit within this limit.

This algorithm allows the model to learn to predict whether the second sentence in a pair naturally follows the first, contributing to its understanding of the flow and structure of text, which is critical for many language processing tasks. The described algorithm prepares the necessary data for the BERT model to be trained on the Next Sentence Prediction (NSP) task. It generates an array of features for each sentence pair in the corpus. These features are as follows:

- **input_ids**: These are the tokenized sequences of two sentences. There's a 50% chance that one sentence will be from a different random document within the corpus. The sentences are tokenized by mapping each word to

an ID based on the pre-trained tokenizer.

- **token_type_ids**: These are a list of binary values (either 0 or 1) that differentiate the two sentences in the input. The first sentence and any padding tokens are marked as 0, and the second sentence is marked as 1. This helps the model distinguish between the two sentences.
- **attention_mask:** This is another list of binary values, where 1 indicates a non-padded token and 0 indicates a padded token. This mask informs the model which tokens should be paid attention to and which should be ignored (the padding).
- **next_sentence_label**: This label indicates whether the second sentence in the pair follows the first sentence in the original document. If it does, the label is 1 ('IsNext'); if the sentence is taken from a different location in the corpus, the label is 0 ('NotNext'). This is the 'ground truth' that the model tries to predict in the NSP task.

Due to the potentially time-consuming nature of the preprocessing stage, the data is saved in a json file after processing. This enables it to be conveniently reloaded for future training sessions, saving computation time. This effective preparation and storage of data streamline the model training process and ensure consistency between training runs.

Chapter 4

Evaluation

The Multiple Choice Question Answering (MCQA) task was chosen as a benchmark for assessing the quality of pre-trained language models due to several reasons.

Firstly, MCQA provides an effective means to evaluate the comprehension capability of language models. This task requires not only understanding the given question but also evaluating multiple possible answers, a process that demands deeper text understanding and reasoning capabilities [39]. It is a challenging task as it goes beyond keyword matching and requires a grasp of nuances and context.

Secondly, MCQA is a practical and common task in many real-world applications. It is not uncommon in fields like education (in tests and examinations), customer service (in automated chatbots), and legal services (in digital legal advisors), to name a few. By choosing MCQA as the benchmark, the evaluation is anchored on a practical and valuable use-case, enhancing the relevance of the models' capabilities.

Lastly, the legal domain adds another layer of complexity. Legal texts are often complicated and nuanced. They require not only a strong understanding of the language but also the underlying legal principles, which are embedded in the word choices and sentence structures. Thus, using a legal MCQA dataset pushes the limits of the language model's understanding abilities.

The collected dataset for this task is a corpus of multiple-choice questions in the legal domain in the Kazakh language. It was manually curated by legal experts and linguists to ensure the questions and answers correctly represent legal terminologies, concepts, and complexities. It provides a good mix of straightforward and complex questions, aiming to cover a broad spectrum of the legal domain. It serves as an effective resource to assess the capability of the pre-trained models in handling the complexities of legal texts in the Kazakh language.

4.1 Multiple Choice Question Answering Dataset

The dataset is a compilation of 2500 multiple-choice questions, each having one correct answer out of four possible options, curated from open-source government quizzes. This dataset serves as a valuable resource for testing knowledge within the legal domain for the civil service of the Republic of Kazakhstan. The collected data covers a broad spectrum of legal topics, reflecting the comprehensive nature of the questions posed in government quizzes. The dataset's construction allows for its direct use in multiple-choice question-answering system training and testing. An example of such a question and its response options is shown in Listing 4.1.

Listing 4.1: Downstream task example question, translated in English.

```
Context: The first session of the newly elected maslikhat is
   ↪ convened by the chairman of the relevant territorial
   ↪ election commission.
A - when there are at least three quarters of the number of
   ↪ deputies designated for this maslikhat
B - when there are at least two-thirds of the number of deputies
   ↪ designated for this maslikhat
C - when there are at least four-fifths of the number of deputies
   ↪  designated for this maslikhat
D - when there is at least one third of the number of deputies
   ↪ designated for this maslikhat
```

```
Ground truth: option A
```

4.1.1 Lexical Analysis of the Kazakh MCQA Dataset

The most frequent unigrams, bigrams, and trigrams in the Kazakh MCQA Dataset highlight important concepts, entities, and themes prevalent in the state employee domain. The analysis of this data sheds light on the specific context of public service and legal norms within the Republic of Kazakhstan.

Unigrams reveal central themes around the state, law, and administrative matters (Figure 1). Entities like 'Kazakhstan' and 'republic' emerge, along with legal terminologies such as 'legal,' 'administrative,' 'normative,' and 'executive.'. It also emphasizes the local context by the terms 'local' and 'central.'

Bigrams accentuate more specific legal contexts such as 'regulatory legal,' 'public services,' and 'anti-corruption.' There's also a reference to the geographic or political structure of the Republic with phrases like 'important city' and 'regional importance.' (Figure 2). Key phrases like 'in accordance with the legislation' underline the rule-bound nature of state functions and public services.

The analysis of trigrams uncovers even more detailed themes and concepts (Figure 3). Phrases like 'providing public services,' 'anti-corruption action,' and 'local executive bodies' denote responsibilities and actions within the state employee domain. 'President of the Republic of Kazakhstan,' 'government of the Republic of Kazakhstan,' and 'state of the Republic of Kazakhstan' emphasize the involvement of different levels of government.

Overall, the frequent unigrams, bigrams, and trigrams in the dataset suggest a strong focus on the legislative and administrative aspect of public services, the geography of political significance, and various levels of governance within the Republic of Kazakhstan.

The majority of questions in the dataset contain a relatively low number of words, indicating a tendency towards concise phrasing in legal MCQA Figure 4.1. More than 800 questions, which constitute almost a third of the total,

have between 1 and 8 words. This shows that the dataset includes a significant number of very short, likely straightforward questions. A slightly smaller number of questions, less than 800, fall in the 8-15 word range. These questions may include a bit more complexity or specificity than the shortest ones, but they still reflect a preference for brevity. More than 400 questions fall in the 15-20 word range, indicating a more moderate level of detail or complexity. These questions constitute a smaller proportion of the dataset, suggesting that such moderately long questions are less common in this legal context. Very long questions, with more than 60 words, are extremely rare, with fewer than 10 instances.

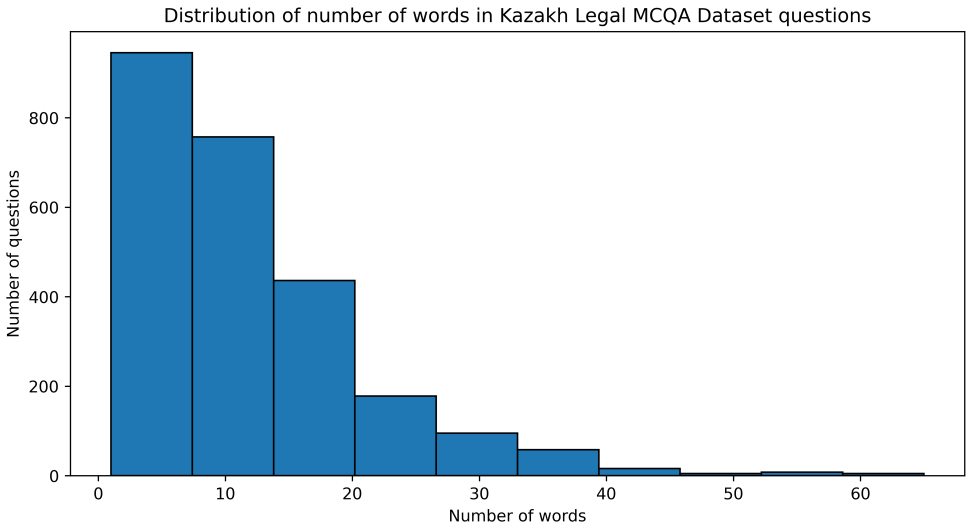

Figure 4.1: Distribution of number of words in Kazakh Legal MCQA Dataset questions

Cosine similiarity The cosine similarity, computed based on text embeddings, essentially measures the semantic similarity between two texts. A high value of cosine similarity indicates a high level of similarity, and vice versa.

In this dataset, a small proportion of questions, less than 30, have a very low similarity with their correct answers (less than 20%) (Figure 4.2). This could suggest that these questions and answers, although logically related, do not share substantial common language or concepts in their text embeddings. In other words, the answers to these questions may be derived from a different context or may be semantically distant from the questions.

A slightly larger proportion, fewer than 500 questions, have a cosine similarity ranging from 20% to 50%. This indicates a moderate level of semantic similarity between these questions and their correct answers.

The majority of questions in the dataset, over 1500, have a higher similarity range from 50% to 75%. This means that for most of the questions, the semantics of the question and its correct answer are fairly similar. This might suggest that the correct answers for these questions are often directly related to the content of the questions themselves.

Lastly, a subset of over 300 questions have a very high similarity with their correct answers (above 75%). This indicates that these questions and their answers share a high level of semantic content. It's possible that these answers might be paraphrasing the questions or using very similar language or concepts.

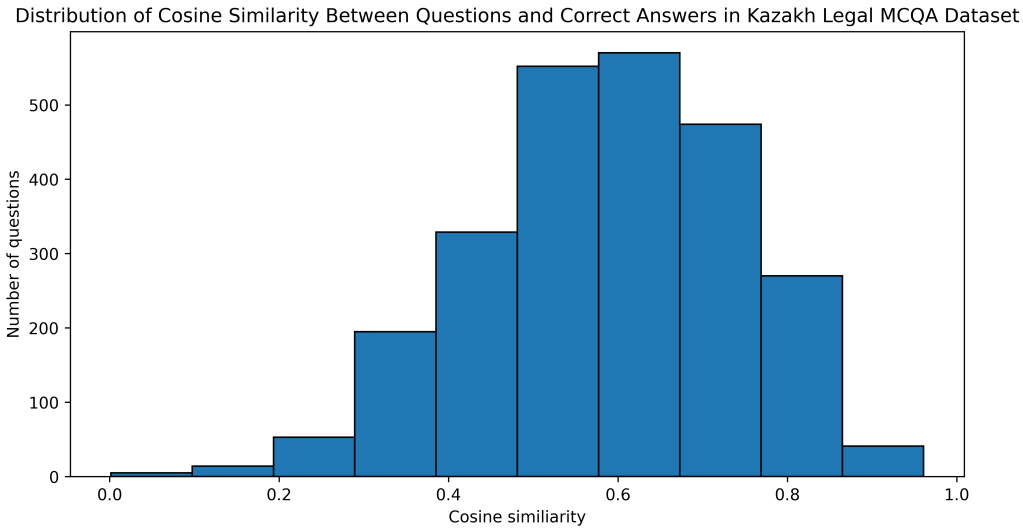

Figure 4.2: Distribution of Cosine Similarity Between Questions and Correct Answers in Kazakh Legal MCQA Dataset based on fasttext embeddings

Distribution of correct answers The bar plot "Distribution of Correct Answers" (Figure 4.3) provides a visual representation of the spread of correct answers across four possible labels in the Kazakh Legal MCQA Dataset.

It is apparent that the dataset exhibits a well-balanced distribution of correct answers. Each of the four possible labels has approximately 600 correct responses. This equal distribution of correct answers across the labels is beneficial in the context of machine learning. When categories are evenly represented,

it is less likely for a model to develop a bias towards a particular label during training.

Such an evenly distributed dataset can contribute to better performance of a model trained on it, as it gets ample opportunity to learn from all classes of data equally. Thus, the model is likely to have a balanced predictive ability for all the answer options.

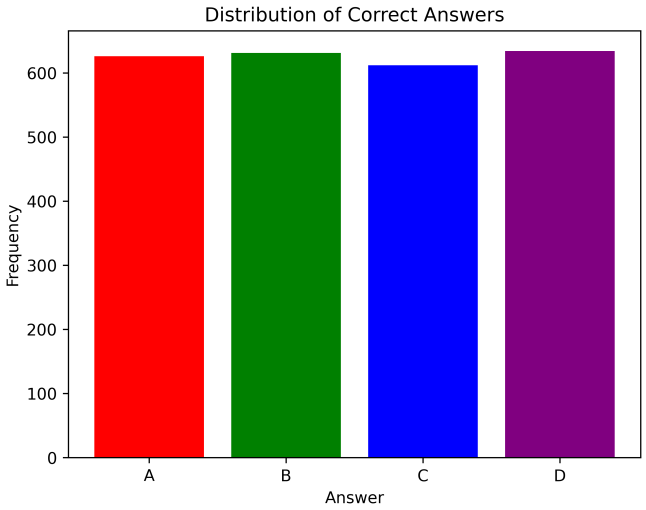

Figure 4.3: Distribution of Correct Answers

4.2 Model configuration

The model configuration for the Multiple Choice Question Answering (MCQA) task leverages the AutoModelForMultipleChoice framework from the Hugging Face Transformers library. This framework is designed to handle multiple-choice tasks, which makes it perfectly suited for this study's requirements.

In terms of hyperparameters, the model configuration is as follows:

1. **Weight decay**: This is set to 0.1 for the AdamW optimizer. Weight decay is a form of regularization that encourages the model to learn smaller weights, effectively reducing complexity and helping to prevent overfitting.

2. **Learning rate**: The learning rate is set to 5e-5. This value controls the step size at each iteration while moving toward a minimum of a loss

function. It plays a crucial role in model convergence and overall training speed.

3. **Batch size**: This is set to 4. The batch size determines the number of training examples utilized in one iteration. Smaller batch sizes can provide a regularizing effect, offering some level of protection against overfitting.

The training process for each selected model was conducted over 10 epochs. An epoch is a complete pass through the entire training dataset. The choice of 10 epochs means each model had an opportunity to learn from each training example ten times over.

The rest of the parameters for the AutoModelForMultipleChoice framework were left at their default settings, adhering to the established configurations provided by Hugging Face's Transformers library. These defaults are often a good starting point, providing a solid foundation for most tasks.

4.3 Dataset encoding for MCQA

The dataset encoding process for Multiple-Choice Question Answering (MCQA) in the work follows these steps:

1. For each of the 2500 questions from the dataset, the text of the question is associated with each of the four answer options. This essentially creates four different 'questions', each consisting of the original question text followed by one of the possible answer choices. Each combination represents a unique instance for the model to consider.

2. After the combinations have been generated, the resulting sequences are tokenized by the pre-trained tokenizer associated with the model being used. Each token is then mapped to an ID based on the model's vocabulary.

3. Following tokenization, the sequences are truncated or padded as necessary to ensure a consistent input length for the model. This is essential as most neural networks, including BERT and its variants, require inputs

of a fixed length.

The dataset encoding process for the Multiple Choice Question Answering (MCQA) task results in an expanded and transformed dataset suitable for training language models.

The goal of this process is to prepare the MCQA data in a format that the model can interpret and learn from. The model is then tasked with determining which of the four sequences (representing the four answer choices) makes the most sense in the context of the question text. This is a standard approach for using transformer-based models, like BERT, for multiple-choice question answering tasks.

From the original 2500 questions, each associated with four possible answer options, the encoding process generates 10,000 unique question-answer combinations. This is because each of the four answer options is paired with the question to form a distinct sequence, effectively amplifying the dataset fourfold. For example question Listing 4.1 will be converted to following 4 sequences Listing 4.2 with corresponding ground truth labels.

For each of these combinations, the encoding process yields the following features:

- **input_ids**: A sequence of token IDs, representing the tokenized form of the question-answer combination. Each unique word is associated with a unique ID according to the model's vocabulary.
- **attention_mask**: A list of integers (0 or 1), where a 1 indicates an active token that the model should pay attention to, and a 0 marks a padded token which should be ignored by the model. This ensures the model only processes the actual content and not the padding used to standardize sequence lengths.
- **label**: A binary marker indicating the correctness of the combination, with 1 signifying the correct answer and 0 for incorrect options. This label serves as the ground truth the model should learn to predict during the training process.

The resulting dataset is then divided into training, validation, and testing subsets. The training set accounts for 70% of the data, which is used to train and adjust the model's weights. The validation set, comprising 15% of the data, is used to tune hyperparameters, prevent overfitting, and assess the model's performance during the training phase. The remaining 15% forms the test set, used to evaluate the model's performance on unseen data, providing an unbiased assessment of its generalization capability.

Listing 4.2: Downstream task example question concatenated answer variants, translated in English.

```
- [CLS] the first session of the newly elected maslikhat is
   ↪ convened by the chairman of the relevant territorial
   ↪ election commission [SEP] with at least three quarters of
   ↪ the number of deputies appointed for this maslikhat [SEP]
- [CLS] the first session of the newly elected maslikhat, the
   ↪ relevant territorial election commission the chairman of
   ↪ the [SEP] convenes the first session of the newly elected
   ↪ maslikhat when at least two-thirds of the number of
   ↪ deputies designated for this maslikhat are present [SEP]
- [CLS] the chairman of the relevant territorial election
   ↪ commission convenes [SEP] when there are at least four-
   ↪ fifths of the number of deputies designated for this
   ↪ maslikhat [SEP]
- [CLS] the first session of the newly elected maslikhat shall be
   ↪  convened by the chairman of the relevant territorial
   ↪ election commission [SEP] with at least one third of the
   ↪ number of deputies designated for this maslikhat [SEP]']
```

4.4 Downstream task results

A rigorous model selection and evaluation process was carried out to measure the performance of the pre-trained models on the multiple-choice question-answering task in the legal domain. The evaluation considered several promi-

nent models, including bert-base-multilingual-cased, bert-base-uncased, xlm-roberta-base, distilbert-base-multilingual-cased, bert-base-multilingual-uncased, as well as variants of BERT models that were pre-trained on a mix of general and legal domain corpora, specifically the Kazakh legal corpus Table 4.1.

Each model underwent fine-tuning over 10 epochs on the multiple-choice question answering task Figure 4.4. The evaluation focused on accuracy, a fundamental metric in machine learning that provides a measure of the overall correctness of the model.

The evaluation yielded the following results:

1. BERT base MLM+NSP pre-trained from scratch on the Kazakh legal corpus outperformed other models, demonstrating the highest accuracy of 56.11%. This result underlines the benefit of domain-specific pre-training, where the model had the opportunity to learn the intricacies of the legal language and specificities of Kazakh.

2. Following closely was the bert-base-multilingual-cased MLM+NSP model pre-trained on the Kazakh legal corpus, achieving an accuracy of 52.12%. The model's performance suggests that incorporating multi-lingual context, even when focused on a particular language (Kazakh, in this case), can still result in substantial performance.

3. Notably, distilbert-base-multilingual-cased showed a competitive performance with an accuracy of 46.27%, indicating that despite its smaller size, DistilBERT managed to encapsulate significant language features that translated to a good performance.

4. The other models, including bert-base-multilingual-uncased, xlm-roberta-base, and bert-base-multilingual-uncased MLM pre-trained on mixed dataset, achieved similar accuracy levels of around 35%. Their performance, while not as high as the top performers, still suggests a reasonable understanding of the Kazakh legal domain, particularly considering the complexity of the task.

5. On the other hand, bert-base-uncased showed the least accuracy at 27.9%,

indicating its inability to effectively understand and process the Kazakh language.

In summary, this evaluation has shown that pre-training language models on a specific domain (here, the Kazakh legal corpus) can significantly enhance their performance on downstream tasks within the same domain. The more tailored the pre-training, the better the performance, particularly when the task involves complex, domain-specific language processing. However, even distilled or generalized models can deliver reasonable results, indicating their potential use in broader applications.

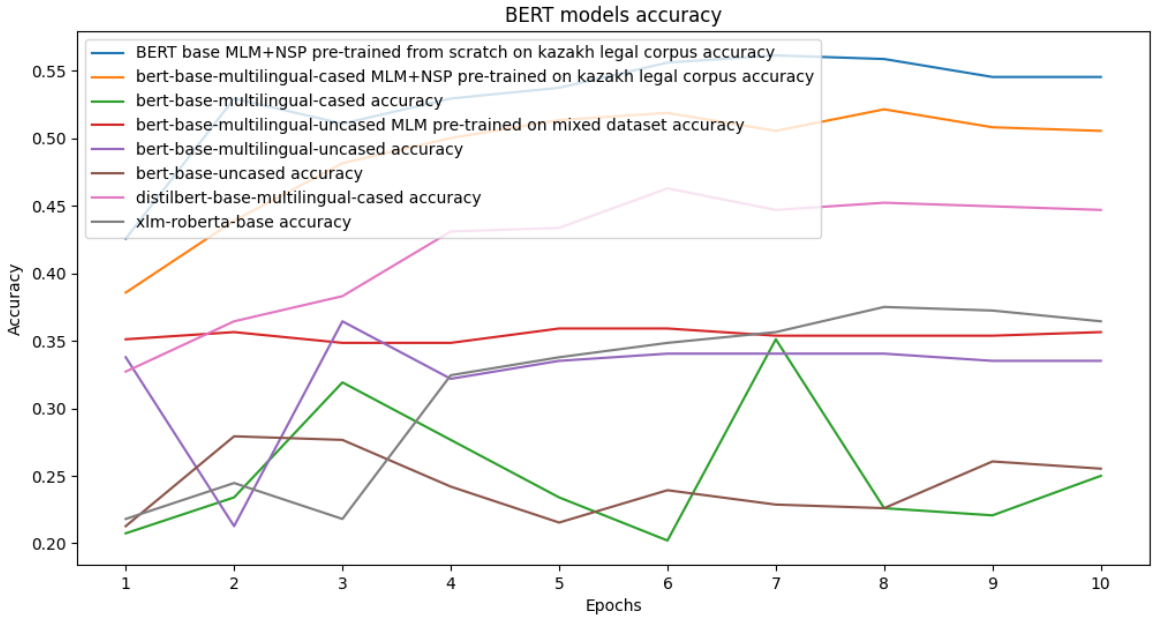

Figure 4.4: Line graphs depicting the evolution of model accuracy over 10 epochs for various fine-tuned BERT models on the Kazakh Legal MCQA Task

4.5 Discussion

The evaluation of several pre-trained language models on the task of multiple-choice question answering in the Kazakh legal domain provided valuable insights into the effectiveness of various approaches. The results strongly suggest that pre-training on domain-specific data improves the performance of models on downstream tasks within that same domain, highlighting the relevance and im-

Model	Accuracy
bert-base-uncased	27.92%
bert-base-multilingual-cased	35.1%
bert-base-multilingual-uncased	36.43%
xlm-roberta-base	37.5%
distilbert-base-multilingual-cased	46.27%
bert-base-multilingual-uncased MLM pre-trained on mixed dataset	35.9%
bert-base-multilingual-cased MLM+NSP pre-trained on kazakh legal corpus	52.12%
BERT base MLM+NSP pre-trained from scratch on kazakh legal corpus	56.11%

Table 4.1: Maximum achieved accuracies over 10 epochs for each fine-tuned BERT-based model on the Kazakh Legal Domain Multiple Choice Question Answering Task

portance of specialized pre-training in the field of Natural Language Processing (NLP).

The BERT model pre-trained from scratch on the Kazakh legal corpus and the bert-base-multilingual-cased model pre-trained on the same corpus outperformed the other models. Their superior performance can be attributed to the domain-specific knowledge they acquired during pre-training. These results demonstrate that even in a multilingual setting, pre-training on specific language data can enhance the model's comprehension of that particular language.

It's noteworthy that the distilled model, distilbert-base-multilingual-cased, despite its smaller size, managed to deliver competitive performance. This finding suggests that distilled models can be efficient alternatives in contexts where computational resources are constrained without drastically compromising the performance.

In contrast, the bert-base-uncased model, which does not understand the Kazakh language, exhibited the least accuracy. This finding corroborates the well-established understanding that language-specific pre-training significantly impacts a model's capacity to process and comprehend that particular language

effectively.

The other models, including xlm-roberta-base, bert-base-multilingual-uncased, and bert-base-multilingual-uncased MLM pre-trained on a mixed dataset, offered performance around the 35% accuracy mark. These models, although not top performers, show that even generalized models can hold a reasonable grasp of domain-specific language, underlining the flexibility and potential of these models in diverse applications.

Overall, the findings underline the potential of language models in complex, domain-specific tasks like legal question answering, pointing to promising avenues for further exploration. Future research can leverage these insights, investigating the enhancement of model performance through further fine-tuning, incorporation of additional legal data for pre-training, or optimization of model architectures to better cater to the peculiarities of the legal language.

The limitations of this study should also be acknowledged. Despite the impressive performance of the models, the accuracies achieved still leave room for improvement. Additionally, while this work focuses on the Kazakh legal domain, different languages and legal systems may present unique challenges that could impact the generalizability of the findings.

This research contributes to the ongoing discourse in the NLP field, particularly in domain-specific language model pre-training. It also has practical implications, given the potential of these models to streamline and enhance the efficiency of legal processes, leading to increased accessibility and understanding of legal procedures and texts for the public.

Chapter 5

Implications and Future Work

5.1 Implications

This research offers several key implications for both academic and practical applications in the Natural Language Processing (NLP) and legal domains, specifically concerning the use of language models in law.

Academic Implications The results of this study contribute significantly to the growing body of research surrounding domain-specific pre-training of language models. It corroborates the hypothesis that pre-training on specific, relevant data sets results in improved performance on downstream tasks within the same domain. The study also points out the efficacy of language-specific pre-training, even in a multilingual setting, enriching the dialogue around the multilinguality aspect of language models.

Furthermore, the noteworthy performance of the distilled model highlights the potential of model distillation in NLP tasks, opening avenues for further research on the development and utilization of lighter, more efficient models without compromising performance drastically.

Practical Implications The practical implications of this study are significant, particularly for the legal industry. By demonstrating that language models can perform reasonably well in domain-specific tasks like legal multiple-choice question answering, this study suggests that these models could be used to enhance the efficiency and accessibility of legal procedures and documents.

Such models could potentially be employed in applications like document review, legal research, contract analysis, and providing guidance to legal practitioners or the general public seeking to understand legal documents or concepts. This could lead to time and cost savings for legal professionals and improved access to legal information for the public.

The success of the distilled model also suggests the feasibility of deploying these models in environments where computational resources are constrained. This could be particularly valuable in settings where access to powerful computational resources is limited but the need for language processing capabilities is high.

Finally, the results of this study may encourage the development of more extensive and diverse corpora for pre-training language models in other under-resourced languages and domains, contributing to the broad goal of democratizing AI.

5.2 Future Work

This research has opened up several promising paths for future work, leveraging the learnings and limitations we encountered during this study.

More Extensive Pre-training One of the key directions for future research will be to undertake a more exhaustive pre-training of BERT and its advanced counterparts on the collected corpus of legal documents. Given the constraints of computing power and time during this research—each of the models required over 30 hours of training on a 10GB GPU—the full potential of extensive pre-training could not be thoroughly explored. Future work could involve employing more advanced computational resources to facilitate comprehensive

pre-training, thereby maximizing the capacity of the model to grasp and apply the intricate semantics and syntactics of the legal language.

Development of Additional Benchmarks Another critical area for future exploration is the development of additional benchmarks to evaluate the efficacy of the language model. While our current study employed the multiple-choice question answering task as the benchmark, this singular metric may not fully capture the model's proficiency in other relevant tasks in the legal domain. Future research could involve devising a more diverse suite of benchmarks that capture various dimensions of legal text comprehension and generation, such as information extraction from contracts or legal question answering. These more diversified benchmarks could provide a more holistic picture of the model's performance in the legal domain.

Exploration of Advanced Language Models Furthermore, given the rapid advancement in language model research, exploring newer and more powerful models, such as GPT-3 or Transformer XL, could be a promising direction. These models could potentially provide better performance and scalability, particularly for complex tasks involving long-term dependencies, a prevalent feature in legal texts.

Overall, our research has laid a solid foundation upon which future work can continue to build, with the ultimate goal of developing highly competent language models tailored for the legal domain.

Chapter 6

Conclusion

Throughout the course of this thesis, significant advancements were made in developing effective language models for legal text processing in the Kazakh language. The thesis began with the assembly of a sizeable corpus of Kazakh legal documents, an important contribution to resources available for NLP research in this language and domain. The compiled corpus, comprised of over 315 million words, represents a rich dataset for training and testing language models.

In addition to the corpus, a Multiple Choice Question Answering dataset was created, containing 2500 questions. This dataset, which is specifically tailored to the civil service examination of the Republic of Kazakhstan, was used as a benchmark for assessing the proficiency of our pre-trained language models in understanding and applying legal concepts and language.

Three language models based on the BERT base architecture were pre-trained, with one of these models being pre-trained entirely from scratch. This process aimed to customize the models to better comprehend and generate Kazakh legal language, going beyond the capabilities of standard, off-the-shelf language models.

We then proceeded to evaluate the effectiveness of eight different models, all variations of BERT, on the multiple-choice question answering task. These experiments were instrumental in understanding the impact of different pre-

training strategies and architectures on the performance of the models.

The findings from these evaluations were instructive. The BERT base model, pre-trained from scratch on the Kazakh legal corpus using both Masked Language Modeling (MLM) and Next Sentence Prediction (NSP) tasks, demonstrated the highest performance. This model achieved an accuracy of 56.11% on the multiple-choice question answering task, underlining the potential of bespoke, pre-training strategies on domain-specific corpora.

This thesis demonstrates that while general-purpose language models provide a good starting point, tailoring models to specific languages and domains through custom pre-training can significantly enhance performance. Going forward, these insights and methodologies can guide similar endeavours in other less-resourced languages and specialized domains. The techniques and results outlined in this thesis form an essential step towards more nuanced and effective AI applications in the legal domain, opening avenues for more efficient and informed decision-making processes.

Bibliography

[1] Nikolaos Aletras, Dimitrios Tsarapatsanis, Daniel Preoţiuc-Pietro, and Vasileios Lampos. Predicting judicial decisions of the European court of human rights: A natural language processing perspective. PeerJ Computer Science, 2016(10):1–19, 2016. ISSN 23765992. doi: 10.7717/peerj-cs.93.

[2] Darkhan Akhmed-Zaki, Madina Mansurova, Gulmira Madiyeva, Nurgali Kadyrbek, and Marzhan Kyrgyzbayeva. Development of the information system for the Kazakh language preprocessing. Cogent Engineering, 8(1), 2021. ISSN 23311916. doi: 10.1080/23311916.2021.1896418.

[3] Zh.D. Mamykova, S. Kumargazhanova, M.R. Aitenova, O.L. Kopnova, V.I. Karyukin, K.M. Barlybay, M. Bolatkhan, and O. Bolatkhan. Development of a Scenario Analysis of the Application of the Information and Analysis System. Bulletin of D. Serikbayev EKTU, (4):139–153, 2022. ISSN 1561-4212. doi: 10.51885/1561-4212_2022_4_139.

[4] Zulfiya Movkebayeva, Dana Khamitova, Aibarsha Zholtayeva, Venera Balmagambetova, and Kairat Balabiyev. Factors influencing the legal regulation and management of education system in Kazakhstan: A review and analysis. Problems and Perspectives in Management, 18(4):14–24, 2020. ISSN 18105467. doi: 10.21511/PPM.18(4).2020.02.

[5] Diana Rakhimova and Aliya Turganbayeva. Lemmatization of big data in the Kazakh language. ACM International Conference Proceeding Series, pages 1–4, 2019. doi: 10.1145/3330431.3330447.

[6] Ualsher Tukeyev, Aliya Turganbayeva, Balzhan Abduali, Diana Rakhimova, Dina Amirova, and Aidana Karibayeva. Lexicon-free stemming

for Kazakh language information retrieval. IEEE 12th International Conference on Application of Information and Communication Technologies, AICT 2018 - Proceedings, pages 3–6, 2018. doi: 10.1109/ICAICT.2018.8747021.

[7] Ilias Chalkidis, Ion Androutsopoulos, and Nikolaos Aletras. Neural legal judgment prediction in English. ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, pages 4317–4323, 2020. doi: 10.18653/v1/p19-1424.

[8] Ilias Chalkidis and Dimitrios Kampas. Deep learning in law: early adaptation and legal word embeddings trained on large corpora. Artificial Intelligence and Law, 27(2):171–198, 2019. ISSN 15728382. doi: 10.1007/s10506-018-9238-9. URL https://doi.org/10.1007/s10506-018-9238-9.

[9] Haoxi Zhong, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. How does NLP benefit legal system: A summary of legal artificial intelligence. Proceedings of the Annual Meeting of the Association for Computational Linguistics, pages 5218–5230, 2020. ISSN 0736587X. doi: 10.18653/v1/2020.acl-main.466.

[10] Michael J. Bommarito II, Daniel Martin Katz, and Eric M. Detterman. LexNLP: Natural language processing and information extraction for legal and regulatory texts. Research Handbook on Big Data Law, pages 216–227, 2021. doi: 10.4337/9781788972826.00017.

[11] Masha Medvedeva, Michel Vols, and Martijn Wieling. Using machine learning to predict decisions of the European Court of Human Rights. Artificial Intelligence and Law, 28(2):237–266, 2020. ISSN 15728382. doi: 10.1007/s10506-019-09255-y.

[12] Haoxi Zhong, Zhipeng Guo, Cunchao Tu, Chaojun Xiao, Zhiyuan Liu, and Maosong Sun. Legal judgment prediction via topological learning. Proceedings of the 2018 Conference on Empirical Methods in Nat-

ural Language Processing, EMNLP 2018, pages 3540–3549, 2018. doi: 10.18653/v1/d18-1390.

[13] Wenmian Yang, Weijia Jia, Xiaojie Zhou, and Yutao Luo. Legal judgment prediction via multi-perspective bi-feedback network. IJCAI International Joint Conference on Artificial Intelligence, 2019-Augus:4085–4091, 2019. ISSN 10450823. doi: 10.24963/ijcai.2019/567.

[14] Daniel Martin Katz, Michael J. Bommarito, and Josh Blackman. A general approach for predicting the behavior of the Supreme Court of the United States. PLoS ONE, 12(4), 2017. ISSN 19326203. doi: 10.1371/journal.pone.0174698.

[15] Aaron Russell Kaufman, Peter Kraft, and Maya Sen. Improving Supreme Court Forecasting Using Boosted Decision Trees. Political Analysis, pages 381–387, 2019. ISSN 14764989. doi: 10.1017/pan.2018.59.

[16] Hai Ye, Xin Jiang, Zhunchen Luo, and Wenhan Chao. Interpretable charge predictions for criminal cases: Learning to generate court views from fact descriptions. NAACL HLT 2018 - 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference, 1:1854–1864, 2018. doi: 10.18653/v1/n18-1168.

[17] Ilias Chalkidis, Manos Fergadiotis, Dimitrios Tsarapatsanis, Nikolaos Aletras, Ion Androutsopoulos, and Prodromos Malakasiotis. Paragraph-level Rationale Extraction through Regularization: A case study on European Court of Human Rights Cases. NAACL-HLT 2021 - 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference, pages 226–241, 2021. doi: 10.18653/v1/2021.naacl-main.22.

[18] L. Karl Branting, Craig Pfeifer, Bradford Brown, Lisa Ferro, John Aberdeen, Brandy Weiss, Mark Pfaff, and Bill Liao. Scalable and explainable legal prediction. Artificial Intelligence and Law, 29(2):213–238, 2021. ISSN 15728382. doi: 10.1007/s10506-020-09273-1.

[19] Josef Valvoda, Tiago Pimentel, Niklas Stoehr, Ryan Cotterell, and Simone Teufel. What About the Precedent: An Information-Theoretic Analysis of Common Law. NAACL-HLT 2021 - 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference, pages 2275–2288, 2021. doi: 10.18653/v1/2021.naacl-main.181.

[20] Ilias Chalkidis, Ion Androutsopoulos, and Achilleas Michos. Obligation and prohibition extraction using hierarchical RNNs. ACL 2018 - 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers), 2:254–259, 2018. doi: 10.18653/v1/p18-2041.

[21] Dan Hendrycks, Collin Burns, Anya Chen, and Spencer Ball. CUAD: An Expert-Annotated NLP Dataset for Legal Contract Review. 2021. URL http://arxiv.org/abs/2103.06268.

[22] Elena Leitner, Georg Rehm, and Julian Moreno-Schneider. Fine-Grained Named Entity Recognition in Legal Documents. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 11702 LNCS:272–287, 2019. ISSN 16113349. doi: 10.1007/978-3-030-33220-4_20.

[23] Jacob Devlin, Ming Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference, 1(Mlm):4171–4186, 2019.

[24] Radford Alec, Wu Jeffrey, Child Rewon, Luan David, Amodei Dario, and Sutskever Ilya. Language Models are Unsupervised Multitask Learners | Enhanced Reader. OpenAI Blog, 1(8):9, 2019. URL https://github.com/codelucas/newspaper.

[25] Mi Young Kim, Ying Xu, Yao Lu, and Randy Goebel. Question answering

of bar exams by paraphrasing and legal text analysis. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 10247 LNAI(July):299–313, 2017. ISSN 16113349. doi: 10.1007/978-3-319-61572-1_20.

[26] Abhilasha Ravichander, Alan Black, Shomir Wilson, Thomas Norton, and Norman Sadeh. Question answering for privacy policies: Combining computational and legal perspectives. EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference, pages 4947–4958, 2019. doi: 10.18653/v1/d19-1500.

[27] Phi Manh Kien, Ha Thanh Nguyen, Ngo Xuan Bach, Vu Tran, Minh Le Nguyen, and Tu Minh Phuong. Answering Legal Questions by Learning Neural Attentive Text Representation. COLING 2020 - 28th International Conference on Computational Linguistics, Proceedings of the Conference, pages 988–998, 2020. doi: 10.18653/v1/2020.coling-main.86.

[28] Kaisar Barlybay. Evaluation of cross-lingual capabilities of BERT models on the example of legal questions with multiple choice answers in the Kazakh language. International Young Scholars Workshop 2020, 2020.

[29] Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. LEGAL-BERT: The muppets straight out of law school. Findings of the Association for Computational Linguistics Findings of ACL: EMNLP 2020, pages 2898–2904, 2020. doi: 10.18653/v1/2020.findings-emnlp.261.

[30] Lucia Zheng, Neel Guha, Brandon R. Anderson, Peter Henderson, and Daniel E. Ho. When does pretraining help?: Assessing self-supervised learning for law and the CaseHOLD dataset of 53,000+ legal holdings. Proceedings of the 18th International Conference on Artificial Intelligence and Law, ICAIL 2021, pages 159–168, 2021. doi: 10.1145/3462757.3466088.

[31] Chaojun Xiao, Xueyu Hu, Zhiyuan Liu, Cunchao Tu, and Maosong Sun. Lawformer: A pre-trained language model for Chinese legal long docu-

ments. AI Open, 2:79–84, 2021. ISSN 26666510. doi: 10.1016/j.aiopen.2021.06.003.

[32] Gaziza Yelibayeva, Altynbek Sharipbay, Gulmira Bekmanova, and Assel Omarbekova. Ontology-based extraction of Kazakh language word combinations in natural language processing. ACM International Conference Proceeding Series, pages 58–59, 2021. doi: 10.1145/3460620.3460631.

[33] E Ondashuly. DIGITALIZATION OF LEGAL SYSTEM OF KAZAKHSTAN ON THE EXAMPLE OF JUDICIAL SYSTEM. JOURNAL OF ACTUAL PROBLEMS OF JURISPRUDENCE, 88:48–52, 2018.

[34] Ellen M. Voorhees and Dawn M. Tice. Building a question answering test collection. SIGIR Forum (ACM Special Interest Group on Information Retrieval), pages 200–207, 2000. ISSN 01635840. doi: 10.1145/345508.345577.

[35] Mi Young Kim, Ying Xu, Yao Lu, and Randy Goebel. Question answering of bar exams by paraphrasing and legal text analysis. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 10247 LNAI:299–313, 2017. ISSN 16113349. doi: 10.1007/978-3-319-61572-1_20.

[36] Ilias Chalkidis, Manos Fergadiotis, Sotiris Kotitsas, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. An empirical study on large-scale multi-label text classification including few and zero-shot labels. EMNLP 2020 - 2020 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference, pages 7503–7515, 2020. doi: 10.18653/v1/2020.emnlp-main.607.

[37] Eneldo Loza Mencía and Johannes Fürnkranz. An evaluation of efficient multilabel classification algorithms for large-scale problems in the legal domain. LWA 2007 - Lernen - Wissen - Adaptivitat - Learning, Knowledge, and Adaptivity, Workshop Proceedings, pages 126–132, 2007.

[38] Lucia Zheng, Neel Guha, Brandon R. Anderson, Peter Henderson, and Daniel E. Ho. When does pretraining help?: Assessing self-supervised

learning for law and the CaseHOLD dataset of 53,000+ legal holdings. Proceedings of the 18th International Conference on Artificial Intelligence and Law, ICAIL 2021, pages 159–168, 2021. doi: 10.1145/3462757.3466088.

[39] Anthony Chen, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. Evaluating question answering evaluation. MRQA@EMNLP 2019 - Proceedings of the 2nd Workshop on Machine Reading for Question Answering, pages 119–124, 2019. doi: 10.18653/v1/d19-5817.

Appendix A

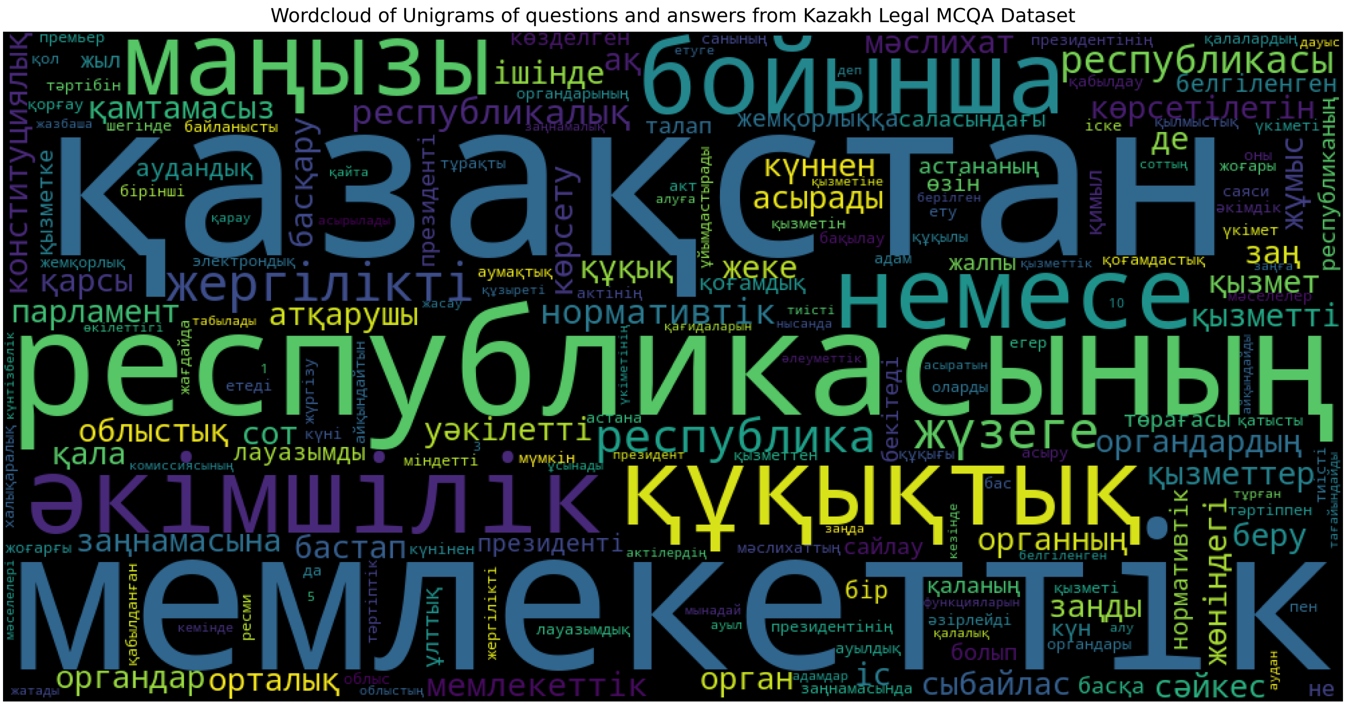

Figure 1: Wordcloud of Unigrams of questions and answers from Kazakh Legal MCQA Dataset

Wordcloud of Bigrams of questions and answers from Kazakh Legal MCQA Dataset

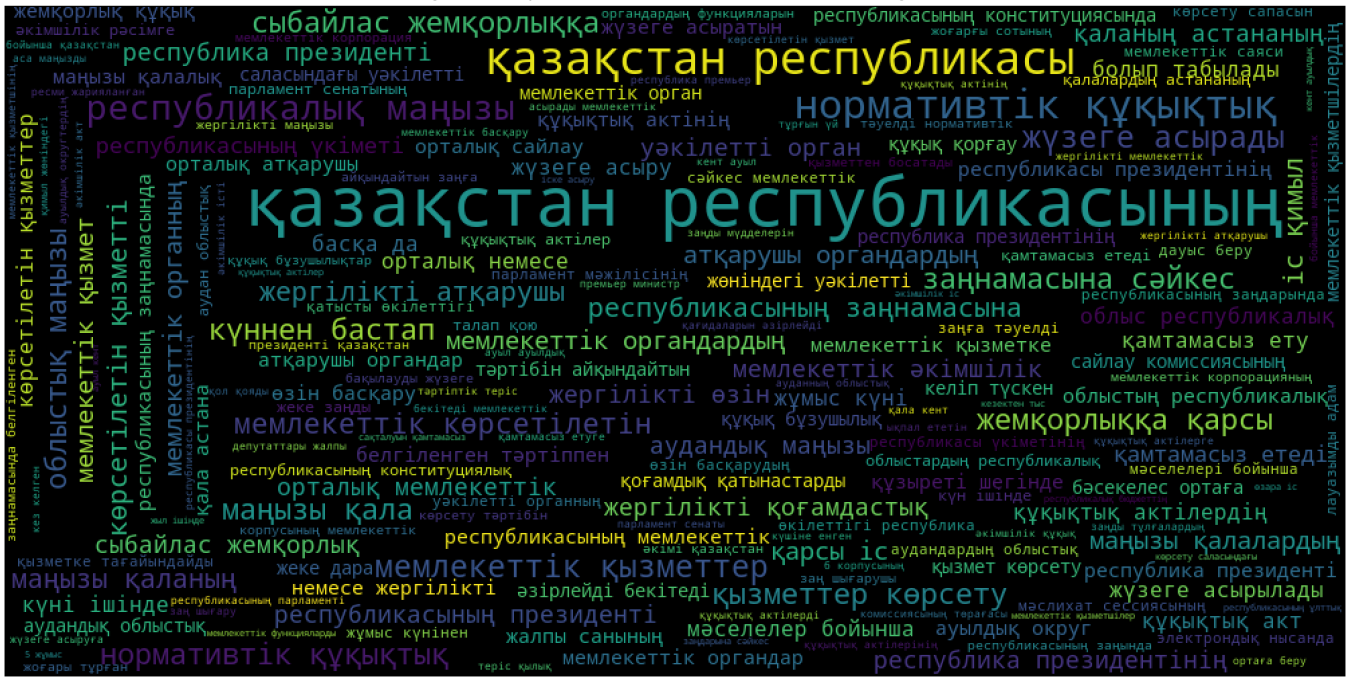

Figure 2: Wordcloud of Bigrams of questions and answers from Kazakh Legal MCQA Dataset

Wordcloud of Trigrams of questions and answers from Kazakh Legal MCQA Dataset

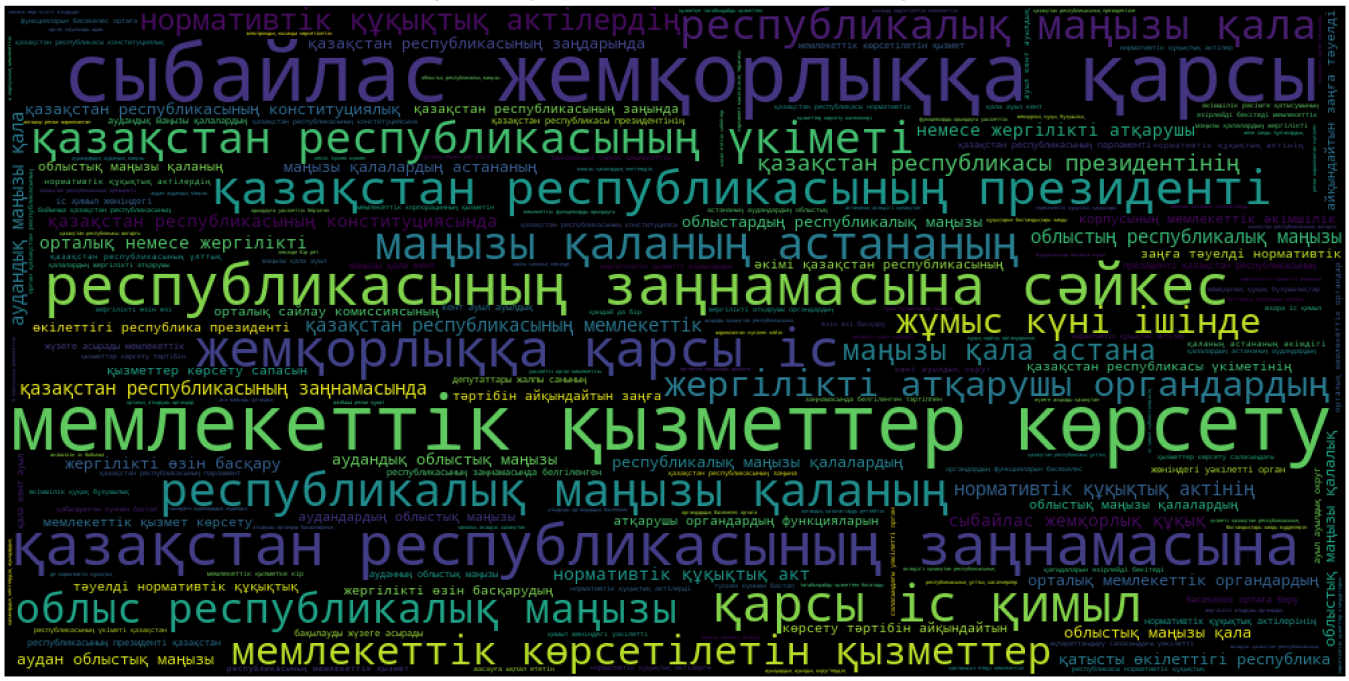

Figure 3: Wordcloud of Trigrams of questions and answers from Kazakh Legal MCQA Dataset

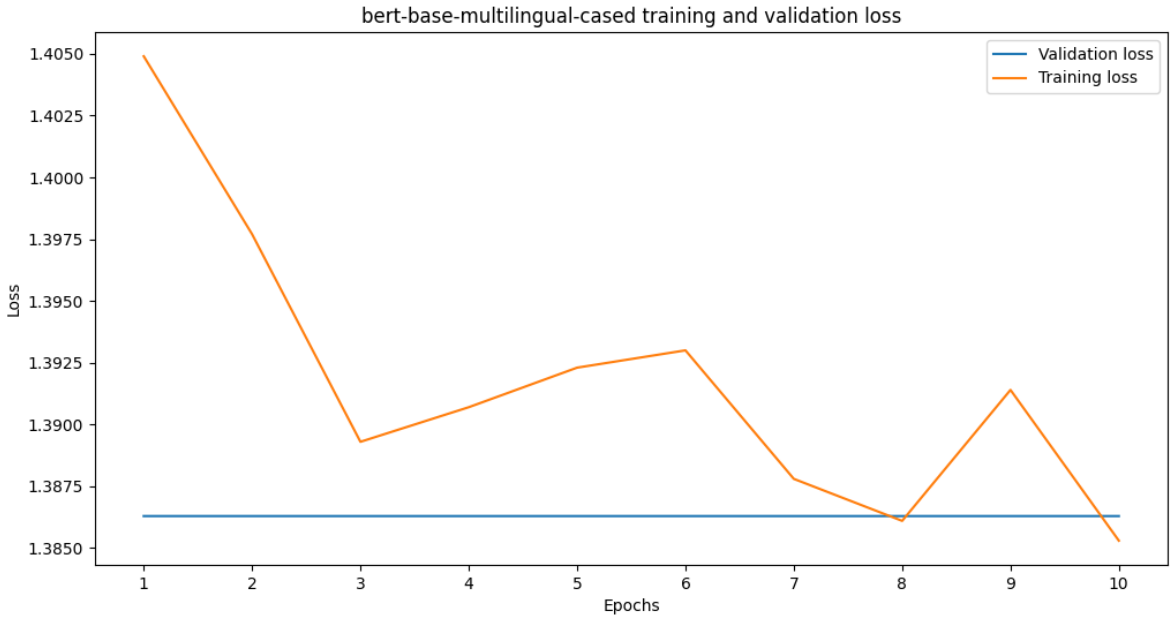

Figure 4: bert-base-multilingual-cased loss

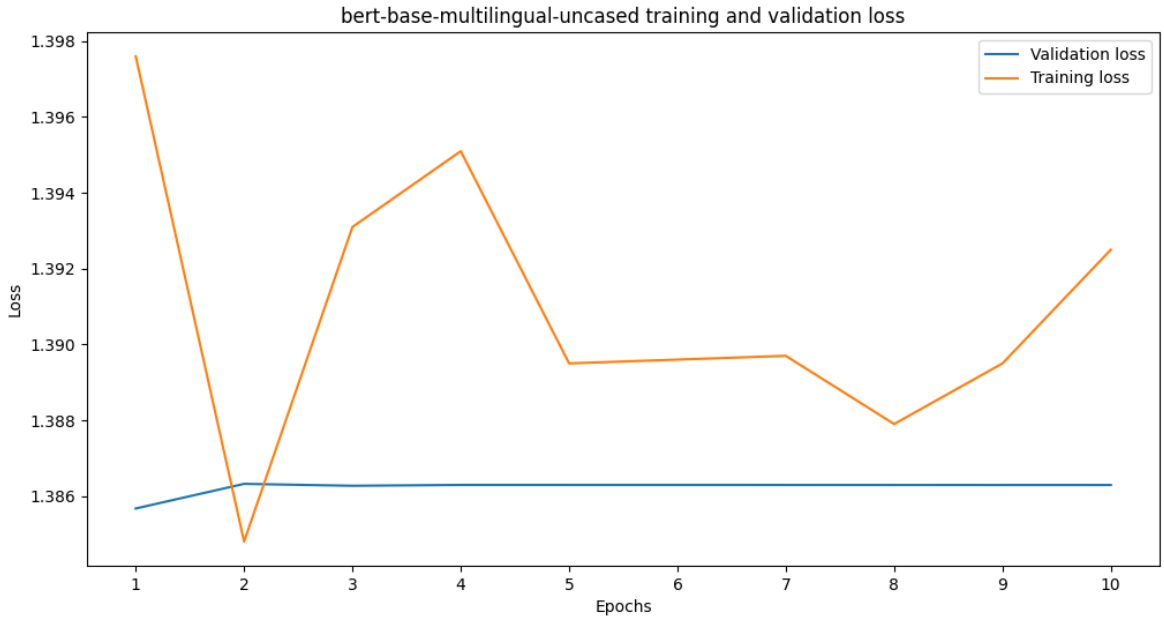

Figure 5: bert-base-multilingual-uncased loss

Figure 6: bert-base-uncased loss

Figure 7: xlm-roberta-base loss

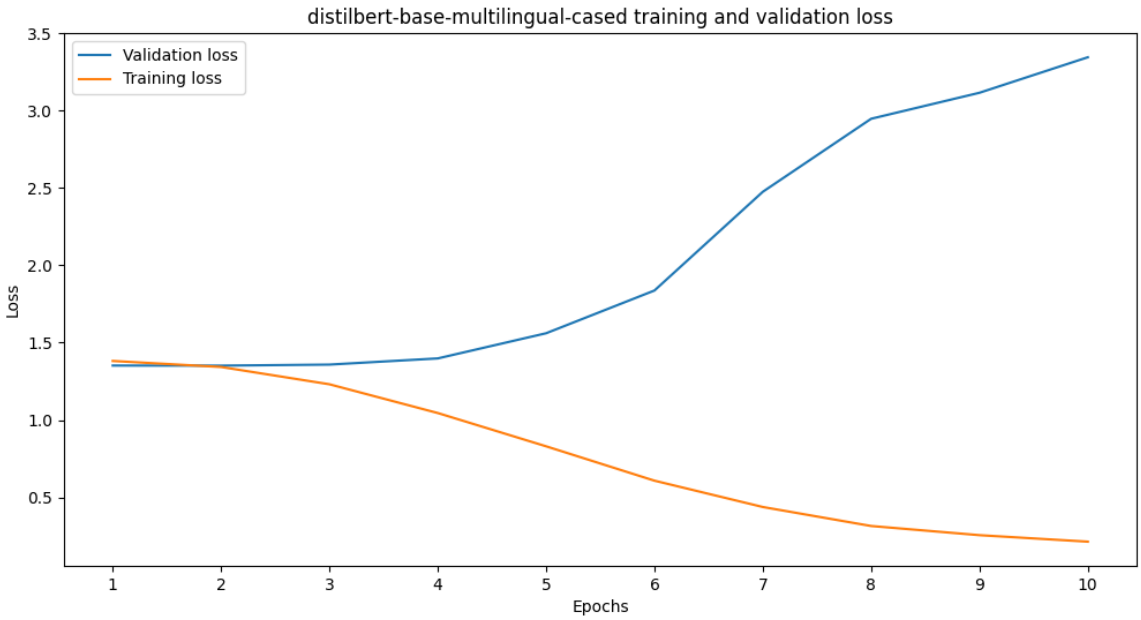

Figure 8: distilbert-base-multilingual-cased loss

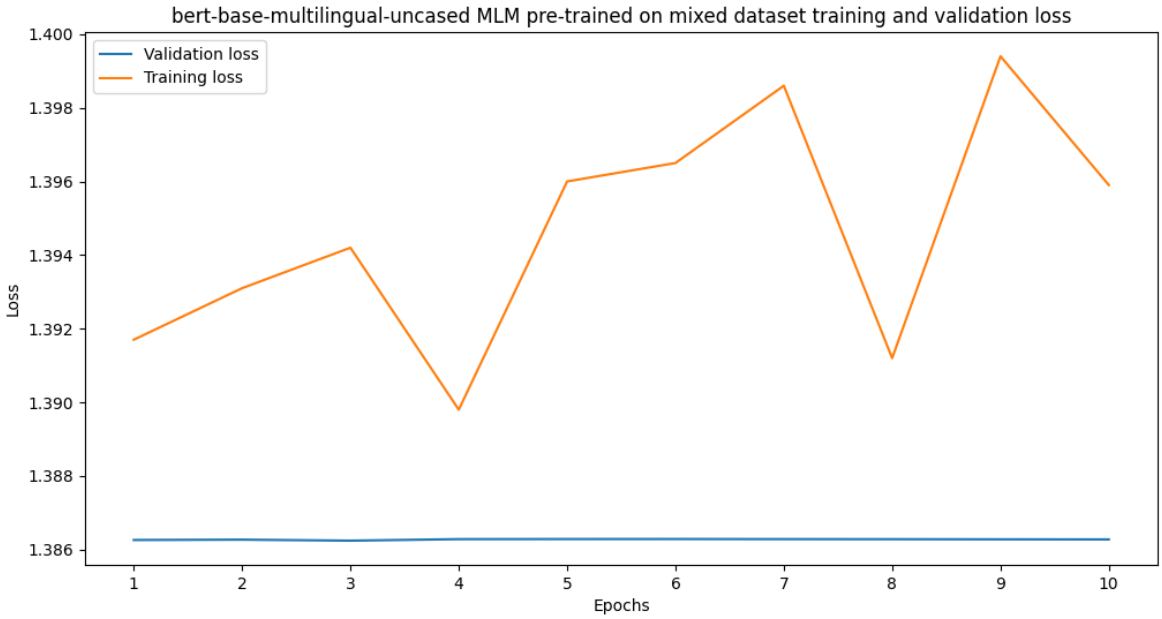

Figure 9: bert-base-multilingual-uncased MLM pre-trained on mixed dataset loss

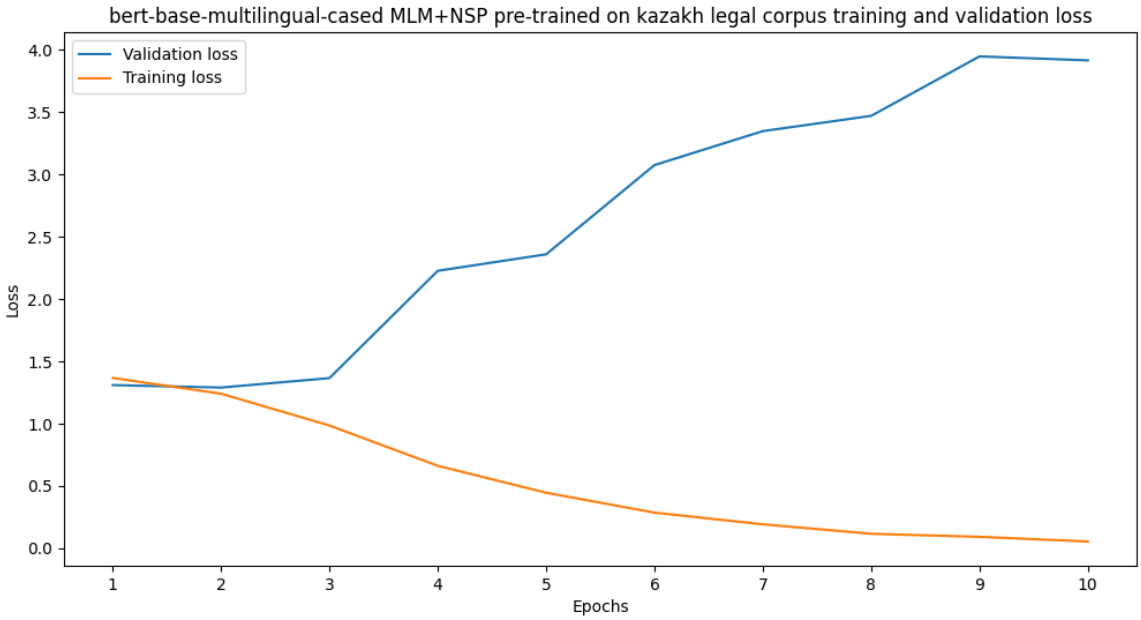

Figure 10: bert-base-multilingual-cased MLM+NSP pre-trained on kazakh legal corpus loss

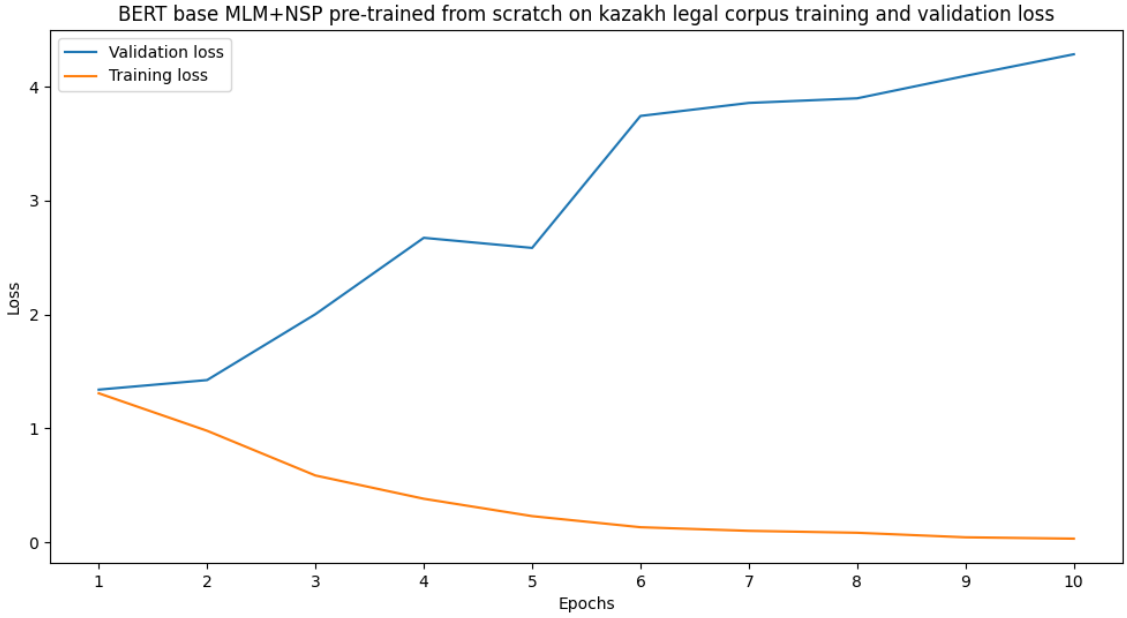

Figure 11: BERT base MLM+NSP pre-trained from scratch on kazakh legal corpus loss