Text Classification of Modern Mongolian Documents using BERT models

Garmaabazar Khaltarkhuu
Graduate School of Information Science and Engineering
Ritsumeikan University
Kusatsu, Shiga, Japan
garmaabazar@gmail.com

Biligsaikhan Batjargal
Research Organization of Science and Technology
Ritsumeikan University
Kusatsu, Shiga, Japan
biligsaikhan@gmail.com

Akira Maeda
College of Information Science and Engineering
Ritsumeikan University
Kusatsu, Shiga, Japan
amaeda@is.ritsumei.ac.jp

Abstract— **This paper investigates the application of state-of-the-art deep-learning-based natural language processing techniques in modern Mongolian documents. In particular, we explore several methods that apply Bidirectional Encoder Representations from Transformers (BERT) models for classifying modern Mongolian legal documents. Based on our findings, we propose BERT-based models called LEGAL-BERT-Mongolian. We demonstrated two variants of LEGAL-BERT-Mongolian, i.e., uncased-LEGAL-BERT-Mongolian and cased-LEGAL-BERT-Mongolian, for classifying modern Mongolian legal documents. The uncased-LEGAL-BERT-Mongolian model achieved the best results, with a precision of 0.91, recall of 0.87, and F1 score of 0.89, whereas the cased-LEGAL-BERT-Mongolian model achieved a precision of 0.87, recall of 0.83, and F1 score of 0.85. Moreover, because the LEGAL-BERT-Mongolian models demonstrated a certain degree of confusion among the "legal," "economy," and "politics" categories, some distinct deep features need to be explored to properly distinguish these classification categories. Furthermore, the proposed LEGAL-BERT-Mongolian models need to be evaluated on larger datasets.**

Keywords— Document classification, Deep learning, Mongolian legal documents, Legal document analysis

I. INTRODUCTION

In organizational management, making decisions at the executive level by analyzing various contracts or legal documents is an important task. There have been increasing demands from researchers, lawyers, executives, and managers to analyze legal documents on a massive scale with prompt and accurate results. Furthermore, a computerized system that makes expert decisions is a long-awaited device for executives and managers who want to avoid problems, conflicts, disputes, and other types of issues. We have an ambitious goal to develop a decision support system (DSS) for helping executives and managers resolve complicated problems and/or questions in a Mongolian context. There has been virtually no deployment of such a system in Mongolia. Specifically, to the best of our knowledge, an approach that applies deep learning techniques to modern Mongolian legal documents has yet to be developed.

A. Problems and Current Situation in Mongolia at the State Level

In his speech at the intermediate evaluation's discussion panel of the "New development" mid-term action plan, the Chief Cabinet Secretary of the government of Mongolia stated that "Over the past 20 years in Mongolia, the government, ministries, and their agencies have issued 517 short or long-term development plans and strategy papers. Currently, 203 of these documents are effective, and many of them overlap or contradict with each other significantly. Only 132 of them are enforced, which has led to less than 26% efficiency" [1]. In addition, according to Worldwide Governance Indicators of the World Bank, the quality of governance in Mongolia as of 2021 is relatively low at 34.62% in effectiveness, 44.71% in regulatory quality, and 46.15% in rule of law [2]. In general, all decisions in Mongolia are highly dependent on the individual decision maker, and many opportunities are lost forever owing to poor and incorrect decisions. It is therefore particularly important to make a professional and prompt decision backed up with decent research using artificial intelligence (AI) systems.

B. Mongolian Language

Mongolian is spoken by the populations in Mongolia and by other ethnic Mongols who live in several provinces of the People's Republic of China and the Russian Federation [3]. Mongolians have used numerous writing systems including a traditional Mongolian script, Square or Phags-pa script, Todo or Clear script, Soyombo script, Horizontal square script, Latin, and Cyrillic script. Besides, the Mongolian language has been written phonetically with Chinese characters [4].

Based on the language reform in 1946, the Cyrillic script was adapted to Mongolian with two additional characters, and Cyrillic became the official Mongolian language script. At that time, the spelling of modern Mongolian in the Cyrillic alphabet was based on the pronunciations in the dialect of the Khalkha, the largest Mongol ethnic group [5] [6]. This was a radical change because traditional Mongolian script preserves the old Mongolian language, whereas modern Mongolian in Cyrillic script reflects the unique pronunciations in modern dialects. Mongolian language is an agglutinative language. In modern Mongolian in Cyrillic script, inflectional suffixes such as plural suffixes, case suffixes, reflexive suffixes, voice suffixes, tense suffixes, aspect suffixes, and mood suffixes are concatenated with the stem. Thus, stemming is necessary for modern Mongolian in Cyrillic script. Moreover, modern Mongolian in Cyrillic script is case-sensitive.

Mongolian legal documents are mainly written in modern Cyrillic script, and this research considers Mongolian legal documents written in modern Mongolian Cyrillic script.

C. Scope of the Present Paper

To achieve our goal of developing a DSS, we investigate how to apply deep-learning-based state-of-the-art natural language processing (NLP) techniques to modern Mongolian legal documents. The current situation and challenges in Mongolia have led us to conduct comprehensive research in

developing a new deep learning model for analyzing modern Mongolian legal documents. In this paper, we report our progress in classifying modern Mongolian legal documents. We believe that text classification is one of the most important tasks in the DSS that we aim to develop.

This paper is organized as follows: In Section II, some related studies are introduced. Text classification tasks of modern Mongolian documents are then introduced in Section III. The proposed LEGAL-BERT-Mongolian models are introduced in Section IV. The experiments and their results are detailed in Section V. Finally, some concluding remarks are given in Section VI.

II. Related Work

A new form of AI, i.e., deep learning, has shown breakthrough results in various fields [7]. It has also achieved a high level of performance across many different NLP tasks in English [8]. State-of-the-art NLP systems for English produce a near-human performance [9].

Within the legal domain, the LawGeex's contract review platform outperforms the 85% accuracy achieved by experienced lawyers on average, with a 94% accuracy rate in reviewing the samples of unseen non-disclosure agreements (NDAs) in English [10]. Chalkidis et al. proposed LEGAL-BERT, which is achieved by fine-tuning the English version of BERT. LEGAL-BERT has been trained on EU, UK, and US legal datasets written in English [11]. Moreover, Chalkidis et al. introduced a collection of datasets, called the Legal General Language Understanding Evaluation (LexGLUE) benchmark, for evaluating the model performance across a diverse set of legal natural language understanding tasks in a standardized manner [12]. Shaghaghian et al. analyzed how different transformer-based language models trained on general-domain corpora can be customized for multiple legal documents reviewing various tasks in English [13].

Caled et al. proposed a hierarchical deep learning model for classifying Portuguese legal documents. They also built a legal corpus of 220K documents written in Portuguese [14]. Using the combination of BERT and Bidirectional Long Short-Term Memory (BiLSTM) models, Wang et al. proposed a classification method for legal questions in China that are written in Mongolian using the traditional Mongolian script [15]. However, their corpus and language model are for the traditional Mongolian script and not for modern Mongolian written in Cyrillic script.

Nevertheless, there has been little research conducted on modern Mongolian NLP, and to the best of our knowledge, no studies have considered AI analyses on modern Mongolian legal documents owing to a lack of research in such areas. We believe that conducting deep-learning-based analyses of modern Mongolian legal documents is an excellent opportunity to demonstrate AI-driven NLP techniques in the legal domain and develop an advanced state-of-the-art system.

As part of its legal system, Mongolia started following civil law in 1992 after abandoning a socialist legal system that had been implemented during the previous 68 years. Most English-speaking countries such as the United States, England, Australia, and Canada have a common law system. The jurisdiction process varies depending on the legal system. We should therefore conduct a detailed analysis for processing modern Mongolian legal documents rather than adopting an as-is language model in English, such as LEGAL-BERT.

III. Text Classification of Modern Mongolian Legal Documents

Because BERT has shown a breakthrough performance in various NLP tasks in English, Erdene-Ochir et al. publicized a Mongolian BERT model pre-trained on 500M Mongolian words extracted from Mongolian news articles with 700M words and a Wikipedia dump as of December 20, 2018 [16]. Among the cased-BERT-Mongolian-base, cased-BERT-Mongolian-large, uncased-BERT-Mongolian-base, and uncased-BERT-Mongolian-large models, in this research, we mainly considered cased-BERT-Mongolian-large and uncased-BERT-Mongolian-large, which have 16 attention heads, 1024 hidden dimensions, and 24 hidden layers with a learning rate of 1e-4, as the original BERT implementation. Mongolian BERT models use Google's SentencePiece [17] with a vocabulary size of 32,000 and a maximum sequence of the first 512 tokens.

A. Existing Text Classification Tasks for Mongolian Language

Gunchinish [18] demonstrated the use of text classification tasks in labeling modern Mongolian news text by fine-tuning pre-trained Mongolian BERT models such as cased-BERT-Mongolian-large and uncased-BERT-Mongolian-large. These tasks were fine-tuned by training 75,661 Mongolian news articles in 9 categories (labels): 1) art and culture, 2) economy, 3) health, 4) legal, 5) politics, 6) sports, 7) technology, 8) education, and 9) nature and environment. The distribution of Mongolian news articles can be seen in Fig. 1. The most frequently appearing news articles in this dataset are in the areas of sports, politics, and economy related news, which account for 17.8%, 17.7%, and 16.4% of all articles, respectively. Among these 75,661 Mongolian news articles, there are 8,285 news articles in the "legal" category, which make up 10.9% of the entire dataset.

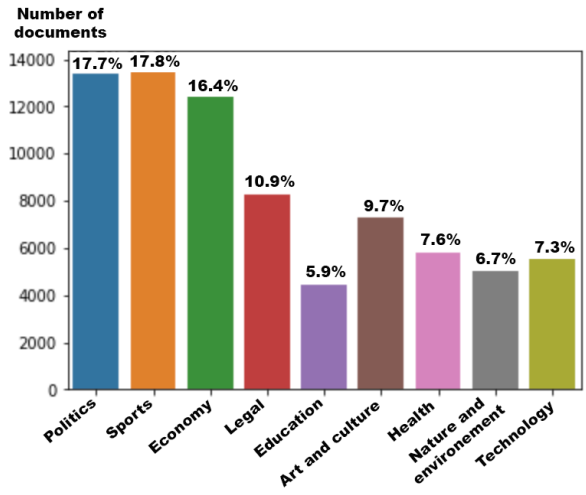

Fig. 1. Numbers of Mongolian news articles in each category.

The performances of these text classification tasks utilizing cased-BERT-Mongolian-large and uncased-BERT-Mongolian-large are shown in Table I and Table II, respectively, based on the precision, recall, and F1 score. The confusion matrices of cased-BERT-Mongolian-large and uncased-BERT-Mongolian-large based text classification tasks are shown in Fig. 2 and Fig. 3, respectively. In these experiments, we split all 75,661 news articles with a training-test split ratio of 80:20 through a random shuffling.

TABLE I. PERFORMANCES IN CLASSIFYING MODERN MONGOLIAN NEWS ARTICLES USING THE CASED-BERT-MONGOLIAN-LARGE.

Category	Precision	Recall	F1 score
Art and culture	0.81	0.70	0.75
Economy	0.57	**0.84**	0.68
Education	0.74	0.44	0.55
Health	0.70	0.76	0.73
Legal	0.67	0.78	0.72
Nature and environment	0.82	0.42	0.55
Politics	0.78	0.78	0.78
Sports	**0.91**	0.78	**0.84**
Technology	0.76	0.64	0.70

TABLE II. PERFORMANCES IN CLASSIFYING MODERN MONGOLIAN NEWS ARTICLES USING THE UNCASED-BERT-MONGOLIAN-LARGE.

Category	Precision	Recall	F1 score
Art and culture	0.89	0.90	0.89
Economy	0.79	0.78	0.79
Education	0.72	0.72	0.72
Health	0.86	0.82	0.82
Legal	0.82	0.82	0.82
Nature and environment	0.71	0.72	0.71
Politics	0.84	0.87	0.85
Sports	**0.95**	**0.95**	**0.95**
Technology	0.87	0.83	0.85

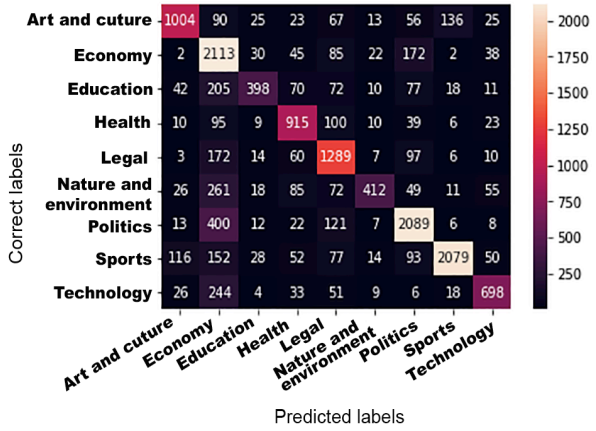

Fig. 2. Confusion matrix of the cased-BERT-Mongolian-large-based text classification task.

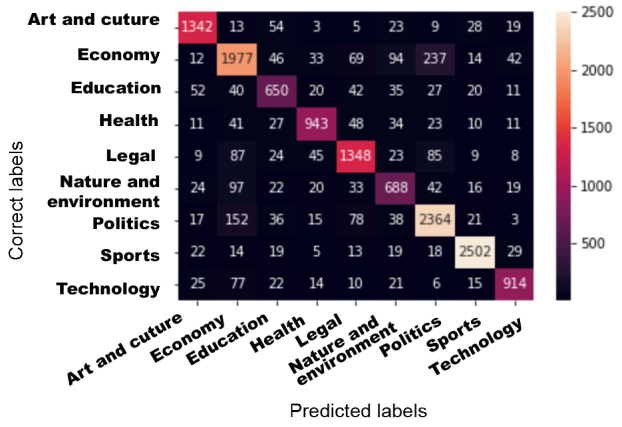

Fig. 3. Confusion matrix of the uncased-BERT-Mongolian-large-based text classification task.

As shown in the above results, the uncased-BERT-Mongolian-large model is capable of labeling sports news more accurately with an F1 score of 0.95, whereas the performance for nature and environment related news was the lowest with an F1 score of 0.71. The F1 score in labeling the "legal" news category was 0.82. The "legal" category of these 75,661 news articles contains not only Mongolian news articles about legal matters but also some parts of official legal documents. In general, uncased-BERT-Mongolian-large achieves a better performance than cased-BERT-Mongolian-large in classifying modern Mongolian news articles. The performance of the BERT-Mongolian-large models in classifying the official Mongolian legal documents is described in the next section.

As shown in Fig. 2 and Fig. 3, in the confusion matrices of the text classification tasks based on cased-BERT-Mongolian-large and uncased-BERT-Mongolian-large, the most confusing categories were "economy" versus "politics," "politics" versus "economy," and "nature and environment" versus "economy." The numbers except the diagonal values in the confusion matrices indicate the numbers of incorrect predictions for each category. The higher numbers correlate to the most confusing categories.

1) Experimental Results in Classifying Modern Mongolian Legal Documents using Existing Text Classification Models.

Further, we conducted experiments to classify modern Mongolian legal documents using the text classification models introduced before. During these experiments, we used the existing fine-tuned text classification models with training batch sizes of 16 and 5 epochs. The 11,323 Mongolian legal documents of mostly unseen data which are described in the next section were presented to the fine-tuned Mongolian BERT models, which were trained on Mongolian news articles. The results of classifying these modern Mongolian legal documents using cased-BERT-Mongolian-large and uncased-BERT-Mongolian-large are shown in Fig. 4 and Fig. 5, respectively.

When the fine-tuned cased-BERT-Mongolian-large model was utilized for classifying modern Mongolian legal documents, 30.5% of the dataset was correctly labeled as "legal." However, 52.2% were labeled as "economy" and 8.8% were labeled as "politics." By contrast, when the fine-tuned uncased-BERT-Mongolian-large model was utilized for classifying modern Mongolian legal documents, 41.8% of the dataset was correctly labeled as "legal." Still, 33.6% were labeled as "politics" and 11.7% were labeled as "economy." In general, uncased-BERT-Mongolian-large shows a better performance than cased-BERT-Mongolian-large in classifying modern Mongolian legal documents. However, existing text classification models become confused among the categories of "legal," "economy," and "politics." Therefore, further investigations into a decent language model for classifying Mongolian legal documents are necessary.

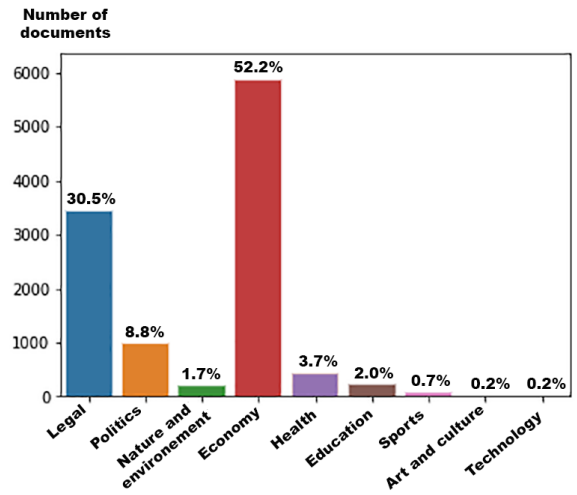

Fig. 4. Classification results of cased-BERT-Mongolian-large-based text classification model over modern Mongolian legal documents.

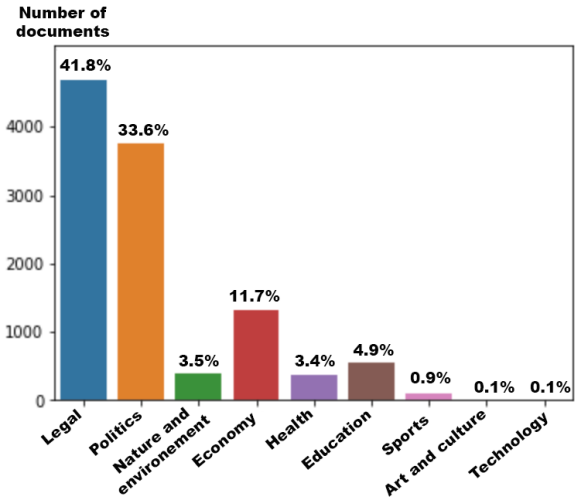

Fig. 5. Classification results of uncased-BERT-Mongolian-large-based text classification model over modern Mongolian legal documents.

IV. LEGAL-BERT-MONGOLIAN: TEXT CLASSIFICATION OF MODERN MONGOLIAN LEGAL DOCUMENTS

We fine-tuned BERT-Mongolian-large models for classifying modern Mongolian legal documents, and named them as LEGAL-BERT-Mongolian. The legal domain is a specialized domain, and a legal text has distinct vocabularies and characteristics. Thus, as described in the previous section, the performance of the existing text classification models in modern Mongolian legal text is unsatisfactory. With LEGAL-BERT-Mongolian, we adapted both cased and uncased versions of BERT-Mongolian-large with additional training on the following 11,323 modern Mongolian legal documents along with 75,661 Mongolian news articles.

A. Modern Mongolian legal documents

Many modern Mongolian legal documents have recently been made publicly available in digital format. However, analyses of these legal documents have not been conducted, mainly owing to a lack of NLP tools that can handle modern Mongolian legal documents. Further computerized analyses are therefore necessary. In this research, the following legal documents including Mongolian laws and decrees of government organizations were prepared for analyzing modern Mongolian legal documents. The legal documents listed in Table III were freely downloaded from the public domain website "Unified Legal Information System" [19] of the National Legal Institute of Mongolia. There are 11,323 modern Mongolian legal documents in 13 legal categories.

TABLE III. UTILIZED MODERN MONGOLIAN LEGAL DOCUMENTS.

Legal category	Number of documents
Resolutions of the government of Mongolia	4,716
Resolutions of the State Ikh Khural (Parliament of Mongolia)	2,066
Resolutions of self-governing bodies (the Citizens' Representative Hurals) of the provinces and capital city Ulaanbaatar	996
Laws of Mongolia	778
Ministerial decrees	748
International treaties concluded or ratified by Mongolia	654
Regulations of government agencies	324
Decisions of the Constitutional Court of Mongolia	295
Decrees of the president of Mongolia	211
Resolutions of the State Supreme Court	174
Decisions of various councils, committees and other collectives	169
Decisions of the heads of the bodies appointed by the Parliament	114
Orders of the governors of the provinces and the mayor of the capital city Ulaanbaatar	78

We fine-tuned LEGAL-BERT-Mongolian models using the above modern Mongolian legal documents. The setup and datasets of LEGAL-BERT-Mongolian are described below.

B. Setup and Datasets

1) Setup

Our model has 16 attention heads, 1,024 hidden dimensions, and 24 hidden layers with a learning rate of 1e-4 and a layer dropout of 0.5. We utilized Google's SentencePiece with a vocabulary size of 32,000 and AutoTokenizer using the maximum sequence of the first 512 tokens, similar to the original BERT implementation. All training was applied with a training batch size of 16 and 5 epochs. There are 337,449,481 parameters in total, 1,843,721 of which are trainable.

2) Datasets

We prepared a dataset by combining 11,323 modern Mongolian legal documents, as introduced before, along with 75,661 Mongolian news articles. The data distribution for each category can be seen in Fig. 6. In this dataset, the "legal" category data containing 1) Mongolian news articles on legal matters, and 2) official Mongolian legal documents were increased to 22.5% of the entire dataset. Sports, politics, and economy related news make up 15.4%, 15.4%, and 14.2% of the dataset, respectively. The classification labels are the same as the categories of the news dataset, i.e., 1) art and culture, 2) economy, 3) health, 4) legal, 5) politics, 6) sports, 7) technology, 8) education, and 9) nature and environment. We split all 86,984 documents with a training-testing split ratio of 80:20 through random shuffling.

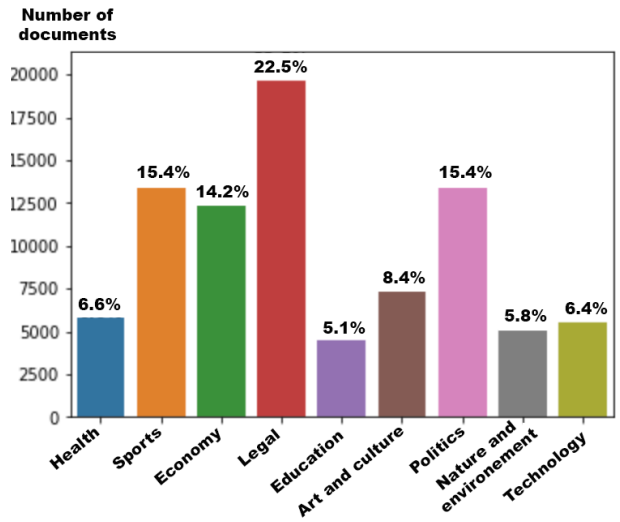

Fig. 6. Data distribution of modern Mongolian documents for each category.

V. Experiments

The performance in classifying modern Mongolian legal documents using the cased and uncased versions of LEGAL-BERT-Mongolian were evaluated by calculating the precision, recall, and F1 score.

A. Experimental Results

The experimental results of classifying modern Mongolian legal documents using cased-LEGAL-BERT-Mongolian and uncased-LEGAL-BERT-Mongolian are shown in Table IV and Table V, respectively, based on the precision, recall, and F1 score. The confusion matrices of cased-LEGAL-BERT-Mongolian and uncased-LEGAL-BERT-Mongolian are shown in Fig. 7 and Fig. 8, respectively.

As shown in Table IV, the performance of cased-LEGAL-BERT-Mongolian model varied. The highest F1 score, 0.85, was achieved for labeling the "legal" category documents. The highest precision was found for labeling the "health" and "nature and environment" category documents at 0.89. The highest recall result of 0.90 was achieved for labeling the "sports" category documents. As shown in Table V, the uncased-LEGAL-BERT-Mongolian achieved a better performance in classifying modern Mongolian legal documents. Although the LEGAL-BERT-Mongolian models were trained on modern Mongolian legal documents, the uncased-LEGAL-BERT-Mongolian model showed the best precision, recall, and F1 score of 0.95 in labeling the "sports" category documents. Moreover, as shown in the confusion matrices of both cased and uncased LEGAL-BERT-Mongolian models shown in Fig. 7 and Fig. 8, respectively, a certain degree of confusion remains in the "legal" category in comparison to the "economy" and "politics" categories.

Because our target is the Mongolian legal domain, we needed to compare the performance of LEGAL-BERT-Mongolian against the existing BERT-Mongolian-large models. In Table VI, we compare the precision, recall, and F1 score of cased-BERT-Mongolian-large, uncased-BERT-Mongolian-large, cased-LEGAL-BERT-Mongolian, and uncased-LEGAL-BERT-Mongolian in classifying modern Mongolian legal documents. The uncased-LEGAL-BERT-Mongolian model shows the best results, with a precision of 0.91, recall of 0.87, and F1 score of 0.89.

TABLE IV. Performances in classifying modern Mongolian documents using the cased-LEGAL-BERT-Mongolian.

Category	Precision	Recall	F1 score
Art and culture	0.68	0.83	0.75
Economy	0.57	0.74	0.66
Education	0.53	0.48	0.51
Health	**0.89**	0.40	0.55
Legal	0.87	0.83	**0.85**
Nature and environment	**0.89**	0.34	0.49
Politics	0.72	0.83	0.77
Sports	0.77	**0.90**	0.83
Technology	0.76	0.57	0.65

TABLE V. Performances in classifying modern Mongolian documents using the uncased-LEGAL-BERT-Mongolian.

Category	Precision	Recall	F1 score
Art and culture	0.88	0.91	0.89
Economy	0.71	0.82	0.76
Education	0.76	0.63	0.69
Health	0.84	0.81	0.82
Legal	0.91	0.87	0.89
Nature and environment	0.79	0.64	0.71
Politics	0.84	0.83	0.84
Sports	**0.95**	**0.95**	**0.95**
Technology	0.78	0.86	0.81

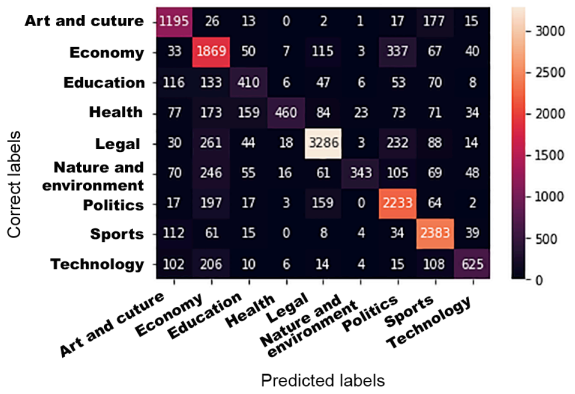

Fig. 7. Confusion matrix of the text classification task of the cased-LEGAL-BERT-Mongolian.

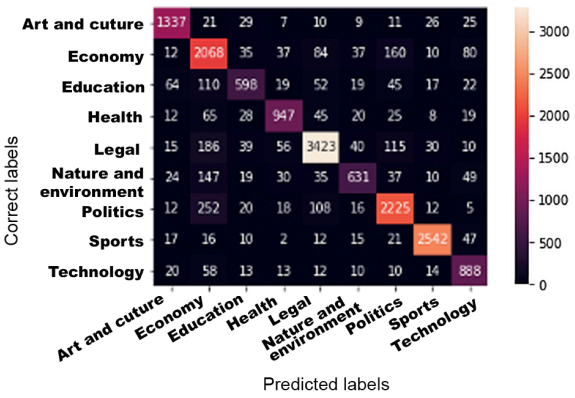

Fig. 8. Confusion matrix of the text classification task of the uncased-LEGAL-BERT-Mongolian.

TABLE VI. Performance comparison in classifying modern Mongolian legaldocuments using various models.

Model	Precision	Recall	F1 score
Cased-BERT-Mongolian-large	0.67	0.78	0.72
Uncased-BERT-Mongolian-large	0.82	0.82	0.82
Cased-LEGAL-BERT-Mongolian	0.87	0.83	0.85
Uncased-LEGAL-BERT-Mongolian	**0.91**	**0.87**	**0.89**

VI. Summary

In this paper, we investigated how to apply existing deep learning models for classifying modern Mongolian legal documents. We proposed a group of BERT-based models called LEGAL-BERT-Mongolian, which we believe are an important part of the DSS that we aim to develop at a future date. The uncased-LEGAL-BERT-Mongolian model achieved the best results with a precision of 0.91, recall of 0.87, and F1 score of 0.89, whereas cased-LEGAL-BERT-Mongolian achieved a precision of 0.87, recall of 0.83, and F1 score of 0.85 in classifying modern Mongolian legal documents. In overall, uncased-BERT models showed a better performance than the cased-BERT models in classifying modern Mongolian legal documents, despite many case-sensitive proper nouns and legal terms in modern Mongolian legal text. In Mongolian text, the first letter of places, organizations, and personal names are always capitalized. Thus, the words written in capital letters have different meanings in a particular context. Legal writing typically requires strict formatting. Likewise, there are some strict rules in the formatting of modern Mongolian legal documents, such as document titles and part/chapter names and their cross references, which should be written in all capital letters.

As a future study, a further detailed error analysis is necessary to investigate how much the capitalized proper nouns and legal terms of Mongolian legal text negatively affect the text classification. Moreover, because the LEGAL-BERT-Mongolian models show a certain degree of confusion between the "legal," "economy," and "politics" categories, some distinct features need to be applied in the proposed models to distinguish the classification categories. Many common or similar sentences may occur in different categories of modern Mongolian text. Another limitation that we need to consider is the token limit of 512 for BERT. Some important parts of Mongolian legal documents might be omitted because the AutoTokenizer utilized in the LEGAL-BERT-Mongolian models automatically truncated texts of longer 512 tokens. Among the various different modern Mongolian legal documents, Mongolian laws have 2,260 words on average, whereas longer laws have many more tokens. Moreover, we should compare the results and classification performance of various neural and non-neural classifiers against LEGAL-BERT-Mongolian. Finally, we also intend to test the proposed LEGAL-BERT-Mongolian model on a dataset of 288K Mongolian court decisions.

Furthermore, we will conduct further research to identify conflicting modern Mongolian legal documents as we believe that it is the another most important tasks in the DSS that we aim to develop.

Acknowledgment

This work was supported in part by JSPS KAKENHI Grant Number 21K12600.

References

[1] The State Great Hural of Mongolia, "Mid-term evaluation's discussion of the medium-term action plan "New development"," http://parliament.mn/n/gico, 2019.

[2] The World Bank Group, "Worldwide Governance Indicators," https://databank.worldbank.org/source/worldwide-governance-indicators, 2021.

[3] J. Hirschberg, and C. D. Manning, "Advances in natural language processing," Science 349(6245), pp. 261–266, 2015.

[4] N. N. Poppe, "Grammar of written Mongolian," University of Washington/Seattle, Washington: Wiesbaden: Otto Harrassowitz, 1954.

[5] T. Shagdarsuren, "Study of Mongolian scripts (Graphic study of grammatology)," National University of Mongolia, Ulan Bator: Urlakh Erdem Khevleliin Gazar, 2001 (in Mongolian).

[6] Q. Sečenbaγatur, B. Tuyaγ-a, and U. Ying, "Mongγul kelen-ü nutuγ-un ayalγun-u sinǰilel-ün uduridqal," Hohhot: Öbür mongγul-un arad-un keblel-ün qoriy-a, 2005 (in Mongolian).

[7] J. Svantesson, A. Tsendina, A. Karlsson, and V. Franzén, "The phonology of Mongolian," New York: Oxford University Press, 2005.

[8] S. Pouyanfar, et al., "A survey on deep learning: Algorithms, techniques, and applications," ACM Computing Surveys (CSUR) 51(5), pp. 1–46, 2018.

[9] T. Young, D. Hazarika, S. Poria, and E. Cambria, "Recent trends in deep learning based natural language processing [Review article]," IEEE Computational Intelligence Magazine 13(3), pp. 55–75, 2018.

[10] LawGeex, "Comparing the Performance of Artificial Intelligence to Human Lawyers in the Review of Standard Business Contracts," https://images.law.com/contrib/content/uploads/documents/397/5408/lawgeex.pdf, 2018.

[11] I. Chalkidis, M. Fergadiotis, P. Malakasiotis, N. Aletras, and I. Androutsopoulos, "LEGAL-BERT: The Muppets straight out of law school," Findings of ACL: EMNLP 2020, 2020.

[12] I. Chalkidis, et al., "LexGLUE: A benchmark dataset for legal language understanding in English," The 60th Annual Meeting of the Association for Computational Linguistics (ACL 2022), Dublin, Ireland, 2022.

[13] S. Shaghaghian, L. Feng, B. Jafarpour, and N. Pogrebnyakov, "Customizing contextualized language models for legal document reviews," In: IEEE International Conference on Big Data (Big Data 2020), pp. 2139–2148, 2020.

[14] D. Caled, M. Won, B. Martins, and M. J. Silva, "A hierarchical label network for multi-label EuroVoc classification of legislative contents," In: Doucet, A., Isaac, A., Golub, K., Aalberg, T., Jatowt, A. (eds.) The 23rd International Conference on Theory and Practice of Digital Libraries (TPDL 2019), LNCS, vol. 11799, pp. 238–252, 2019.

[15] G. Wang, F. Bao, and W. Wang, "Mongolian questions classification in the law domain," In: International Conference on Asian Language Processing (IALP 2020), pp. 56–59, 2020.

[16] T. Erdene-Ochir, Sh. Gunchinish, and E. Bataa, "BERT pretrained models on Mongolian datasets," https://github.com/tugstugi/mongolian-bert/.

[17] Google, "SentencePiece," https://github.com/google/sentencepiece.

[18] Sh. Gunchinish, "Mongolian text classification," https://github.com/sharavsambuu/mongolian-text-classification.

[19] Unified Legal Information System, https://legalinfo.mn/en.