COMPUTER SCIENCES AND INFORMATION TECHNOLOGIES

UDC 004.855.5(045)
DOI:10.18372/1990-5548.81.18977

[1]V. M. Sineglazov,
[2]I. M. Savenko

LANGUAGE MODEL ADAPTATION FOR LEGAL UKRAINIAN DOMAIN

[1]Aviation Computer-Integrated Complexes Department, Faculty of Air Navigation Electronics and Telecommunications, National Aviation University, Kyiv, Ukraine
[2]Department of Artificial Intelligence, Institute for Applied System Analysis, National Technical University of Ukraine "Sikorsky Kyiv Polytechnic Institute," Kyiv, Ukraine
E-mails: [1]svm@nau.edu.ua ORCID 0000-0002-3297-9060, [2]savenko.ilya@lll.kpi.ua

Abstract**—***Language models in recent decades make a huge step towards solving the tasks that previously could be done only by humans. Development of NLP area is different scopes gives an opportunity to solve domain specific tasks and transfer knowledge from learnt data towards the useful inferences based on that. This article provides the NLP model approach in specific legal domain. Additionally, this article explores performance of pre-training small models and its utilization and checks the scores on fine-tuned task of checking sentence similarities via SBERT. According to this articles it is proven that domain-specific pre-trained models can perform better results than generally trained language model. This article also provides the language model that is adopted to the Ukrainian legal domain.*

Index Terms—Intellectual text analysis; natural language processing; text embeddings; opinion mining; machine learning; BERT; SBERT; Legal-BERT.

I. INTRODUCTION

Domain-specific language reflects the syntactic and semantic representation of the language style that could be used in specific areas. While the general-purpose language model is learned from the text with a common style (Wikipedia, fiction books [8]) the performance of the model could be degraded on down-stream tasks tight to some areas. Hence, to level up the productivity of the model, it is needed to use a different approach. Some could come up with using different data to train the model or change the structure of the model for better performance. This article reflects the first approach – pre-training on the documents with domain-specific language to align the weights.

Previously this task was done by representing the text via statistical representation and wrappings [14], [15]. The brightest one is Word2Vec [18], [19] and GloVe [15], [16], [17]. But the performance of the models was constrained with opportunities to solve simple tasks [20].

This article provides a comprehensive analysis of domain-specific languages (DSLs) tailored to the Ukrainian language, specifically focusing on the application of Bidirectional Encoder Representations from Transformers (BERT [19]) in the analysis of legal texts. By examining the syntactic and semantic characteristics of the Ukrainian legal language, this study investigates how DSLs and BERT [19] models are constructed and adapted to effectively handle these linguistic nuances. The research encompasses the development and implementation of Ukrainian-specific DSLs and BERT [19] models, their performance in parsing and interpreting legal documents, and their potential contributions to legal informatics. This analysis aims to enhance our understanding of language-specific computational models and their efficiency in representing and processing Ukrainian legal texts using advanced NLP techniques.

The interest of the research appears in the author due to the low quantity of articles in the NLP area tight to the Ukrainian language and tasks suitable in the specific domain. But the process of research is moving on. This process involves also model generation [1], [2], generating NERs [3] to fine-tuning existing common-dictionary models [4].

In scope of solving tasks in domain-specific language, the major part is done in English language. Particularly, for the legal documents the Legal-BERT [5] model was trained to produce more sophisticated results. In the article references above there is also comparison report created for the check of performance models between models trained on common text corpuses and pre-trained text corpuses regarding the US law documents. Furthermore, there is a works that build process of finding conflict identifications (Aires [6]). Also one paper should be noted for presenting usage of BERT [19] for legal textual entailment prediction (Wehnert [7]).

This article focuses on SBERT implementation and comparison analysis of text from the Unified state register of court decisions. Based on the data from the state register BERT model is trained.

II. Related Works

To initialize the discussion about domain-specific adaptation it is wise to start from the source model BERT Base [19]. This is a pre-trained language model based on some general domain that includes the fiction literature and Wikipedia articles. The route approach of pre-training the model is masking technique. It describes the build of prediction of the next sequences based on the text data that includes into model training. With some parameters it is tuned to identify the general text sequences occurrences based on the previous ones. As a result of using the general text corpuses the model has a low performance in scope of its using in domains.

The Sci-BERT [10] model and some others [12], [13] proves that using different text-corpuses dataset that is more specific to the language domain that is going to be analyzed gives more sophisticated result in down-stream tasks regardless the architecture of the model.

As this article focuses only on legal domain it is reasonable to make more specific description of the legal scope. Previously, the legal domain adaptation is described in article of Legal-BERT [5] model. The dataset of pre-trained model includes over 450000 documents from US and EU data. The specification of the data includes more complex level of language: the specific vocabulary and stylistics, hence the parameters of the describing model is suited for the data. The following adjustments are done for such specification: learning rates are used in both variant – lower and higher, dropout rates – crucial factor when dealing with diverse legal texts, batch sizing and training epochs – more complex variations are explored. As a result, domain adaptation gave more elaborate result in legal down-stream tasks. The loss distribution reflects that training model from scratch gives the best results of model performance. The middle case – is further pre-trained model based on BERT-Base. The worst results are fetched from the general model.

Almost the same was explores in the JuriBERT [11] article. The article focuses on small BERT models and French language. This article outlines the optimization of the size of domain-specific models. The models trained on some specific dataset and optimized training hyperparameters that have less size could also give the qualitative results in scope of narrow task.

III. Problem Statement

A. *Pre-training task*

We present a mathematical model of machine learning for natural language processing. For the pre-training from scratch we will be using a Masked Language Modeling approach. Given the sequence of tokens:

$$X = (x_1, x_2, \ldots, x_n).$$

The approach includes the random replacing some tokens with [MASK] token resulting in a modified sequence $\tilde{X}$. The goal for the sub-problem is to find mapping with minimized loss score for predicting the original token x_i from the masked sequence $\tilde{X}$. The function could be defined in such way:

$$Loss_{MLM} = -\sum\nolimits_{i\in M} \log\log p\left(x_i|\tilde{X}\right),$$

where M is the set of masked positions and $p\left(x_i|\tilde{X}\right)$ is the probability of the correct token x_i given the model's softmax output at position *i*.

The other sub-problem is to set up next sentence prediction. This involves predicting whether a sentence *B* is the actual next sentence that follows *A* in the original text. The definition could be described in such way: having a pair of sentences (*A*, *B*). The model outputs is a binary label $y \in \{0,1\}$ of indicator function. “1” stands for approval of *B* following *A* and “0” – vice versa. The loss functions could be defined in such way:

$$Loss_{NSP} = -\left[y \log p + (1-y)\log(1-p)\right],$$

where *p* is the model predicted probability that *B* follows *A*.

General problem that is reflected in final hidden state corresponding to the first input token [CLS] is used as the aggregate sequence representation for classification task. If *C* is the number of classes, and h_{CLS} is the hidden state then predicted value could be described in such manner:

$$p = \text{softmax}\left(Wh_{CLS} + b\right),$$

where *W* and *b* is the trainable weights of the network. The general loss function could be described in such manner:

$$Loss_{class} = -\sum_{k=1}^{C} y_k \log(p_k),$$

where y_k is the indicator value for class *k*, and p_k is the predicted probability for class *k*.

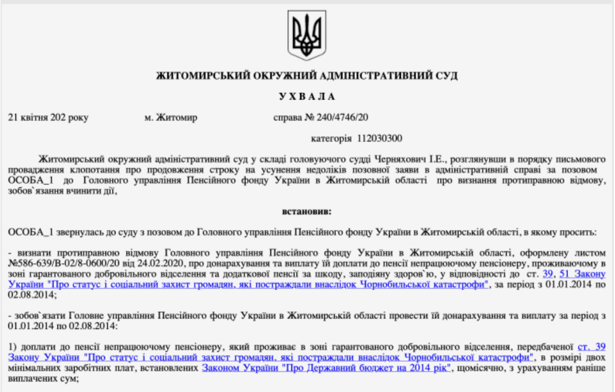
B. Down-stream task

In scope of this articles the down-stream task took into consideration to analyze the performance of the language model. As a downstream task we explore the SBERT [9] paragraph analysis to count the closeness of each of them to its corresponding to the section of the Ukrainian legislation. This article represents checks of closeness cross intersected between the text chunks. The main goal is to define how precisely the text chunk wrappings would be from the same part of legislation. As the SBERT model is used for such problem the necessity to do complex further fine-tuning is absent. The Sentence-BERT is based on the original BERT.

IV. Model Description

In scope of the task of this article we take the data from the database of legal document of the court of Ukraine. The documents that we took for training purposes was the court decision. While making a quick glance at the document we can make a conclusion that it’s structure consists of three parts: The heading, main subject part of the court case and conclusions. The quantity of the documents that was took for training was 50000 documents. The documents were choose randomly regardless the time and scope of the court processing event.

This was done to check if it is feasible for the model to track the elaborated semantic nuances of legal abstraction represented in the document. The research regarding this paper stopped on the section identification level abstraction (Fig. 1).



ЖИТОМИРСЬКИЙ ОКРУЖНИЙ АДМІНІСТРАТИВНИЙ СУД

У Х В А Л А

21 квітня 202 року м. Житомир справа № 240/4746/20

категорія 112030300

Житомирський окружний адміністративний суд у складі головуючого судді Черняхович І.Е., розглянувши в порядку письмового провадження клопотання про продовження строку на усунення недоліків позовної заяви в адміністративній справі за позовом ОСОБА_1 до Головного управління Пенсійного фонду України в Житомирській області про визнання протиправною відмову, зобов`язання вчинити дії,

встановив:

ОСОБА_1 звернулась до суду з позовом до Головного управління Пенсійного фонду України в Житомирській області, в якому просить:

- визнати протиправною відмову Головного управління Пенсійного фонду України в Житомирській області, оформлену листом №586-639/В-02/8-0600/20 від 24.02.2020, про донарахування та виплату їй доплати до пенсії непрацюючому пенсіонеру, проживаючому в зоні гарантованого добровільного відселення та додаткової пенсії за шкоду, заподіяну здоров`ю, у відповідності до ст. 39, 51 Закону України "Про статус і соціальний захист громадян, які постраждали внаслідок Чорнобильської катастрофи", за період з 01.01.2014 по 02.08.2014;

- зобов`язати Головне управління Пенсійного фонду України в Житомирській області провести їй донарахування та виплату за період з 01.01.2014 по 02.08.2014:

1) доплати до пенсії непрацюючому пенсіонеру, який проживає в зоні гарантованого добровільного відселення, передбаченої ст. 39 Закону України "Про статус і соціальний захист громадян, які постраждали внаслідок Чорнобильської катастрофи", в розмірі двох мінімальних заробітних плат, встановлених Законом України "Про Державний бюджет на 2014 рік", щомісячно, з урахуванням раніше виплачених сум;

Fig. 1. Example picture of the court decision

For BERT pre-training we took the standart masking properties. It is 15% masking of the word sequences. For the pre-training we took Bert-Tiny and Bert-Mini models.

This Table I represents the sizes of the model that will be taken into consideration.

Table I Sizes of the Model

Model	**Architecture**	**Params**
BERT tiny	$L = 2,\ H = 128,\ A = 2$	6 M
BERT mini	$L = 4,\ H = 256,\ A = 4$	15 M
BERT small	$L = 8,\ H = 512,\ A = 8$	42 M
BERT base	$L = 12,\ H = 768,\ A = 12$	110 M

As it seen from the Table I the model was trained on major 4 models. Starting from the BERT tiny model is represented with 2 layers, 128 hidden units and 2 attention heads. BERT mini model has 4 layers, 256 hidden units and 4 attention heads. Next, BERT small model has 8 layers, 512 hidden units and 8 attention heads. Last, but not least we took BERT base model with 12 layers and attention heads and 768 hidden units. In the next section will describe how stepping into this list with such parameters will give the improvement in result of semantic closeness.

The other part of the model training is hyperparameters set-up. We found out that best optimal configuration is to use 2e-4 learning rate step. The batch size for the training is used in size of 8. We also take 15 train epochs to see the scale of the error function descent behavior and then get the conclusions in this paragraph.

Final point to take into consideration is environment where the training was performing. The software service that was used for such purpose is Google Colab. For the model training purposes the hardware that was taken is A100 GPU.

V. Results

For comparison reason we took several models to check. As the universe of the pre-training models is quite small, we took multilingual RoBERTa model with extracted weights for the part of Ukrainian and Russian language. The further comparison of the paragraph wrappings give the result that could be interpreted as overtraining, hence we will not take them into the consideration of conscious(all wrappings was presented in narrow scope between 0.98 and 0.99).

Based on pre-trained models we make a measure check how the chunks of text curpuses are presented in vector space. Such cases were took into consideration:

1) Text from the administrative violation section of the court case

2) Text from the other section of the law violation.

3) Random Ukrainian fact reflected in text

4) Random Ukrainian text from fiction book.

After implementing SBERT approach we got such result (Table II).

TABLE II THE RESULT WAS OBTAINED AFTER IMPLEMENTING THE SBERT APPROACH

Model	**Administrative section court case**	**Other violation court case**	**Neutral-language fact**	**Text from fiction book**
BERT tiny	0.6383	0.4735	0.4453	0.3624
BERT mini	0.7344	0.5132	0.4612	0.3855
BERT small	0.8114	0.6448	0.5094	0.4613
BERT base	0.9382	0.6743	0.379	0.2435

From the table given above we can conclude that depending on the size of the model that we trained, as the size is bigger – the more precise semantic text representation we have. This results shown that using SBERT and build even small models they can catch the context representations of the domain-specific language.

VI. CONCLUSIONS

This paper was aimed to decompose the problem of language adaptation of huge corpuses of texts in scope of legal domain. During the research we build the model that could be used for downstream tasks. All research was performed on the data of Ukrainian legal text. After the pre-training we found out that pre-trained model can define the closeness of resulting vectors of random text taken from different part of legal-violation sections of the database by counting its cosine similarity. This article shows that paragraphs of the text from the same section of the legal-violation database semantically more close to each other than paragraphs from different sections.

On the contrary, we also took the random Ukrainian text that legal database doesn't consists of and do the same closeness operation check. The result of this check shows that random Ukrainian text has lower score of closeness than even the paragraph parts from different legal database sections.

With the well-known techniques of NLP we showed that pre-trained model with even low quantity of entries that was used during the training could solve outstanding problems and build the semantic representation of some specific domain. We showed this on the example of legal domain that is complex to operate.

The future of this research work stream is to go deeper into the Ukrainian legal domain. Based on this articles we can check further how not only parts of the text gives semantic text representation but also create the knowledge graph to extract the more sensible chunks.

REFERENCES

[1] Stefan Fischer, Kateryna Haidarzhyi, Jörg Knappen, Olha Polishchuk, Yuliya Stodolinska, and Elke Teich, "A Contemporary News Corpus of Ukrainian (CNC-UA): Compilation, Annotation, Publication," *In Proceedings of the Third Ukrainian Natural Language Processing Workshop (UNLP) @ LREC-COLING*, 2024, pp. 1–7, Torino, Italia. ELRA and ICCL.

[2] Maria Shvedova and Arsenii Lukashevskyi, "Creating Parallel Corpora for Ukrainian: A German-Ukrainian Parallel Corpus (ParaRook||DE-UK)," *In Proceedings of the Third Ukrainian Natural Language Processing Workshop (UNLP) @ LREC-COLING*, 2024, pp. 14–22, Torino, Italia. ELRA and ICCL.

[3] Dmytro Chaplynskyi and Mariana Romanyshyn, "Introducing NER-UK 2.0: A Rich Corpus of Named Entities for Ukrainian," *In Proceedings of the Third Ukrainian Natural Language Processing Workshop (UNLP) @ LREC-COLING*, 2024, pp. 23–29, Torino, Italia. ELRA and ICCL.

[4] Artur Kiulian, Anton Polishko, Mykola Khandoga, Oryna Chubych, Jack Connor, Raghav Ravishankar, and Adarsh Shirawalmath, "From Bytes to Borsch: Fine-Tuning Gemma and Mistral for the Ukrainian Language Representation," *In Proceedings of the Third Ukrainian Natural Language Processing Workshop (UNLP) @ LREC-COLING*, 2024, pp. 83–94, Torino, Italia. ELRA and ICCL.

[5] Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos, "LEGAL-BERT: The Muppets straight out of Law School," *In Findings of the Association for Computational Linguistics: EMNLP*, 2020, pp. 2898–2904, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.findings-emnlp.261

[6] J. P. Aires, D. Pinheiro, V. S. d. Lima, et al., "Norm conflict identification in contracts," *Artif Intell Law*, 25, 397–428, 2017. https://doi.org/10.1007/s10506-017-9205-x

[7] S. Wehnert, S. Dureja, L. Kutty, et al., "Applying BERT Embeddings to Predict Legal Textual

Entailment," *Rev Socionetwork Strat*, 16, 197–219, 2022. https://doi.org/10.1007/s12626-022-00101-3

[8] Christopher Manning, Prabhakar Raghavan, and Hinrich Schütze, "Introduction to Information Retrieval," *Cambridge University Press*, 2008. https://doi.org/10.1017/CBO9780511809071.

[9] Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David Duvenaud, "Neural Ordinary Differential Equations," *32nd Conference on Neural Information Processing Systems (NeurIPS 2018*), Montréal, Canada, 2018, arXiv preprint arXiv:1908.10084. Retrieved from https://arxiv.org/abs/1908.10084

[10] I. Beltagy, K. Lo, & A. Cohan, "SciBERT: A Pretrained Language Model for Scientific Text," *Computer Science > Computation and Language*, 2019, arXiv preprint arXiv:1903.10676. Retrieved from https://arxiv.org/abs/1903.10676

[11] I. Chalkidis, M. Fergadiotis, P. Malakasiotis, N. Aletras, & I. Androutsopoulos, "JuriBERT: A Masked-Language Model Adaptation for French Legal Text" 2020, *arXiv preprint arXiv:2110.01485*. Retrieved from https://arxiv.org/abs/2110.01485

[12] Y. Ganin, & V. Lempitsky, "Unsupervised domain adaptation by backpropagation," In *Proceedings of the International Conference on Machine Learning (ICML)*, 2015, pp. 1180–1189.

[13] M. Wang, & W. Deng, "Deep visual domain adaptation: A survey," *Neurocomputing,* 312, pp. 135–153, 2018. https://doi.org/10.1016/j.neucom.2018.05.083

[14] Tomáš Mikolov, *Statistical language models based on neural networks*, Ph.D. thesis, Brno University of Technology, 2012.

[15] Tomáš Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean, *Efficient estimation of word representations in vector space*. arXiv:1301.3781 [cs], January 2013.

[16] Jeffrey Pennington, Richard Socher, and Christopher Manning, "GloVe: global vectors for word representation," *In Proc. of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543, Doha, Qatar. Association for Computational Linguistics, October 2014. https://doi.org/10.3115/v1/D14-1162.

[17] Jeffrey Pennington, Richard Socher, and Christopher Manning, "GloVe: global vectors for word representation," *In Proc. of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543, Doha, Qatar. Association for Computational Linguistics, October 2014. https://doi.org/10.3115/v1/D14-1162.

[18] T. T. Vu, V. A. Nguyen, & T. B. Le, "Combining Word2Vec and TF-IDF with Supervised Learning for Short Text Classification," *In 2020 3rd International Conference on Computational Intelligence (ICCI)*, 2020, pp. 241–245.

[19] J. Devlin, M. W. Chang, K. Lee, & K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, vol. 1, (Long and Short Papers) (pp. 4171–4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

[20] M. Lin, S. Liao, & Y. Huang, "Hybrid word2vec and TF-IDF approach for sentiment classification," *Journal of Information Science*, 45(6), 797–806, 2019.

Received May 17, 2024.

Sineglazov Victor. ORCID 0000-0002-3297-9060. Doctor of Engineering Science. Professor. Head of the Department of Aviation Computer-Integrated Complexes.
Faculty of Air Navigation Electronics and Telecommunications, National Aviation University, Kyiv, Ukraine.
Education: Kyiv Polytechnic Institute, Kyiv, Ukraine, (1973).
Research area: Air Navigation, Air Traffic Control, Identification of Complex Systems, Wind/Solar power plant, artificial intelligence.
Publications: more than 700 papers.
E-mail: svm@nau.edu.ua

Savenko Illia. Post-graduate Student.
Artificial Intelligence Department, Institute for Applied System Analysis, National Technical University of Ukraine "Ihor Sikorsky Kyiv Polytechnic Institute", Kyiv, Ukraine, (2023).
Research Interests: artificial neural networks, artificial intelligence, programming.
Publications: 2.
E-mail: savenko.ilya@lll.kpi.ua

В. М. Синєглазов, І. М. Савенко. Порівняльний аналіз методів векторизації тексту
В роботі розглянуто способи векторизації текстових властивостей природної мови в контексті задачі інтелектуального аналізу тексту. Проаналізовано найпоширеніші способи статистичного аналізу вилучення ознак та методи з урахуванням контексту. Проведено опис вищезазначених типів обрамлення тексту та їх найпоширеніші реалізації. Виконано їх порівняльний аналіз, який показав зв'язок між типом задачі

інтелектуального аналізу тексту та методом, що показує найкращі метрики. Описано та реалізовано топологію нейронної мережі, яка стоїть в основі вирішення задачі та отримання метрик. Порівняльний аналіз проведено за допомогою відносного аналізу часу теорії алгоритмів та метрик класифікації: accuracy, f1-score, precision, recall. Метрики класифікації узято з результатів побудови моделі нейронної мережі з використанням описаних методів обрамлення. В результаті в задачі аналізу тональності тексту найкращим виявився статистичний метод обрамлення на основі n-грамів символьних послідовностей.

Ключові слова: інтелектуальний аналіз тексту; обробка природної мови; обрамлення тексту; аналіз думок; машинне навчання; BERT; SBERT; Legal-BERT.

Синєглазов Віктор Михайлович. ORCID 0000-0002-3297-9060.
Доктор технічних наук. Професор. Завідувач кафедри авіаційних комп’ютерно-інтегрованих комплексів. Факультет аеронавігації, електроніки і телекомунікацій, Національний авіаційний університет, Київ, Україна.
Освіта: Київський політехнічний інститут, Київ, Україна, (1973).
Напрям наукової діяльності: аеронавігація, управління повітряним рухом, ідентифікація складних систем, вітроенергетичні установки, штучний інтелект.
Кількість публікацій: більше 700 наукових робіт.
E-mail: svm@nau.edu.ua

Савенко Ілля Михайлович. Аспірант.
Кафедра штучного інтелекту, Інститут прикладного системного аналізу, Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського», Київ, Україна.
Освіта: Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського», Київ, Україна, (2023).
Напрям наукової діяльності: штучний інтелект, машинне навчання, штучні нейронні мережі, програмування.
Кількість публікацій: 2.
E-mail: savenko.ilya@lll.kpi.ua