Training Specialised Chatbots on Ukrainian Scientific Text Corpora Using Transfer Learning

Roman O. Liashenko
Kryvyi Rih State Pedagogical University
Kryvyi Rih, Ukraine
ORCID iD 0009-0000-2614-6997

Serhiy O. Semerikov
Kryvyi Rih State Pedagogical University, Kryvyi Rih, Ukraine
Institute for Digitalisation of Education of the NAES of Ukraine
Kyiv, Ukraine
Zhytomyr Polytechnic State University, Zhytomyr, Ukraine
Kryvyi Rih National University, Kryvyi Rih, Ukraine
Academy of Cognitive and Natural Sciences, Kryvyi Rih, Ukraine
ORCID iD 0000-0003-0789-0272

***Abstract*—This paper presents an extensive experimental study on applying transfer learning techniques to fine-tune pre-trained language models for creating chatbots specialised in Ukrainian scientific texts. The study involves curating two diverse text corpora from research publications, selecting appropriate base models, and fine-tuning them on the prepared datasets using the Hugging Face transformers library. The fine-tuned models are tested for generating chatbot responses to domain-specific user queries. Detailed analyses of the datasets, model architectures, fine-tuning process, and evaluation results are provided. The results demonstrate the feasibility and effectiveness of transfer learning for adapting language models to Ukrainian scientific texts and creating chatbots capable of meaningful dialogue in specialised domains. Challenges and future research directions are discussed.**

***Index Terms*—chatbots, transfer learning, language models, fine-tuning, Ukrainian language, scientific texts.**

I. Introduction

The rapid development of artificial intelligence and natural language processing technologies has sparked growing interest in creating software agents capable of engaging in natural language dialogue, known as chatbots. Leading IT companies like Google, Microsoft, Meta, and OpenAI are actively working on chatbot development [1]. Successful projects like ChatGPT from OpenAI demonstrate the significant potential of such systems in various spheres of human activity.

Chatbots as natural language AI systems have immense potential to enhance the efficiency and quality of various human activities by automating routine processes, providing intelligent user support, and enabling personalised learning. Successful developments in this field can fundamentally change how humans interact with computer systems, increasing work productivity and learning effectiveness. However, addressing ethical and security issues is crucial for implementing these technologies in society's best interests.

While chatbots have shown promising results in general domains, their application to specialised areas like scientific texts in low-resource languages such as Ukrainian remains underexplored. This paper aims to investigate the effectiveness of transfer learning techniques for adapting pre-trained language models to Ukrainian scientific text corpora and creating chatbots capable of domain-specific dialogue.

II. Overview of Chatbot Training Approaches

A. Supervised Learning

Supervised learning is one of the main approaches to training chatbots, involving training models on labelled "query-response" pairs. Architectures based on recurrent neural networks (e.g., Seq2Seq with LSTM) and transformers (e.g., GPT) are used for this purpose [2]. These models can generate contextually relevant responses but require large amounts of high-quality labelled data.

Seq2Seq models consist of an encoder that processes the input sequence (user query) and generates its vector representation, and a decoder that uses this representation to generate the output sequence (chatbot response). The key feature of Seq2Seq models is the use of recurrent layers, particularly LSTM, for processing sequential data. However, they can suffer from the vanishing gradient problem when processing long sequences [3].

Transformer architectures, especially GPT models, represent a state-of-the-art approach to building language models and dialogue systems. They outperform other language models like recurrent neural networks due to their self-attention mechanism and ability to process sequences of arbitrary length while maintaining context [4].

B. Reinforcement Learning

Reinforcement learning allows models to learn through interaction with an environment (user) and receive feedback in the form of rewards. This approach is implemented using generative adversarial networks (GAN) and the iterative process of reinforcement learning based on human feedback (RLHF) [5].

RLHF typically involves iteratively training a reward model on human judgments, optimizing a policy model to maximize the predicted reward, collecting new data using the updated policy, and retraining the reward model on the expanded dataset [6].

979-8-3315-4263-4/24/$31.00 ©2024 IEEE

GANs are an innovative approach to dialogue management in chatbots. They consist of two neural networks—a generator and a discriminator—that compete with each other during training. The generator aims to create responses indistinguishable from human-generated ones, while the discriminator learns to distinguish between generated and real responses.

Tran et al. [7] proposed a model combining reinforcement learning and GANs to generate both accurate and human-like responses. The RL component acts as a backbone, with the discriminator from the adversarial learning module evaluating the relevance of RL-generated responses. The discriminator output serves as an additional reward for the RL generator, encouraging the model to generate responses indistinguishable from human ones.

C. Transfer Learning

Transfer learning involves using knowledge gained by a model in solving one task to improve its performance in solving another similar task. The most common approach is fine-tuning a pre-trained model on a specific dataset to adapt it to a particular domain [6].

Fine-tuning existing large language models instead of creating them from scratch is often more practical and efficient due to resource and data efficiency, knowledge transfer, high performance, lower entry barrier, continuous learning, and broad applicability [8].

The fine-tuning process includes loading pre-trained model parameters, preparing a task-specific dataset, extracting features using pre-trained layers, adjusting model parameters via backpropagation and gradient descent with a lower learning rate, updating gradients and applying regularization to prevent overfitting, selecting a fine-tuning strategy (full or partial model tuning), and evaluating performance and optimizing hyperparameters (Figure 1).

During fine-tuning, changes occur at the architecture and individual neuron levels, including weight adjustment, activation function output changes, top layer tuning with lower layers frozen, feature space adjustment for the new task, last layer replacement for task correspondence, and batch normalization parameter updates [6]. Fine-tuning allows effectively adapting pre-trained models to specific tasks while preserving their "intuition" and optimizing for the new task's specifics.

D. Chatbot Evaluation Metrics

To evaluate the effectiveness of chatbot training, both automatic metrics and human evaluation methods are used, allowing for the assessment of the relevance, coherence, and naturalness of generated responses. Common automatic metrics include BERTScore, perplexity, BLEU, and ROUGE [9]. Table I provides an overview of these metrics.

Human evaluation methods involve direct assessment of chatbot performance by real users or experts, providing a more comprehensive evaluation of model effectiveness, considering aspects such as relevance, coherence, and usefulness of generated responses.

TABLE I
METRICS FOR EVALUATING TEXT GENERATION MODELS [9].

Metric	Description
BERTScore	Compares generated text with reference text using BERT embeddings
Perplexity	Evaluates how well the probability distribution predicted by the model matches the actual data distribution
BLEU	Calculates n-gram precision scores between generated and reference text
ROUGE	Calculates n-gram recall scores between generated and reference text

III. METHODOLOGY

A. Dataset Creation

Two datasets were created for chatbot training, containing texts of scientific publications in the field of information technology.

The first dataset, "CEUR Workshop Proceedings", includes materials from scientific conferences and seminars covering a wide range of research in computer science and engineering. The dataset was created by downloading the https://ceur-ws.org/ website, extracting texts from 68,791 PDF files corresponding to volumes 1-3583 for 1995-2024, and creating a single 1917 MB text file. The relative distribution of publications by article language according to Scopus is: English – 94.807%, Russian – 1.368%, German – 1.101%, Spanish – 0.772%, Portuguese – 0.691%, Turkish – 0.596%, French – 0.340%, Ukrainian – 0.160%, Italian – 0.131%, Czech – 0.015%, other languages combined – 0.019%.

The second dataset, "Information Technologies and Learning Tools", includes articles on theoretical and applied aspects of using information and communication technologies in education. The dataset was created by downloading the journal's website https://journal.iitta.gov.ua/index.php/itlt, extracting texts from 1,732 PDF files corresponding to volumes 1-100 for 2006-2024, and creating a 107 MB text file. The approximate relative distribution of publications by article language according to Web of Science is: Ukrainian – 52.54%, Russian – 26.73%, English – 20.73%.

The created datasets differ in volume, time coverage, thematic focus, and distribution of writing languages, allowing for a comparative analysis of the effectiveness of chatbot models trained on heterogeneous text corpora.

B. Model Selection and Fine-Tuning

Considering the prevalence of GPT models, the possibilities of accessing their modern versions (GPT 3.5, 4.0) and alternative models (Gemini 1.0, Claude 3) were studied. However, due to the unpredictable cost of fine-tuning and using these models, the historical GPT-2 model [10], trained on OpenAI's internal WebText dataset (40 GB), was chosen instead.

For fine-tuning on the "CEUR Workshop Proceedings" dataset, the GPT2-XL model was selected, considering its commensurability with the WebText dataset and the predominance of English texts. For the "Information Technologies and Learning Tools" dataset, the gpt2-uk model, pre-trained

979-8-3315-4263-4/24/$31.00 ©2024 IEEE

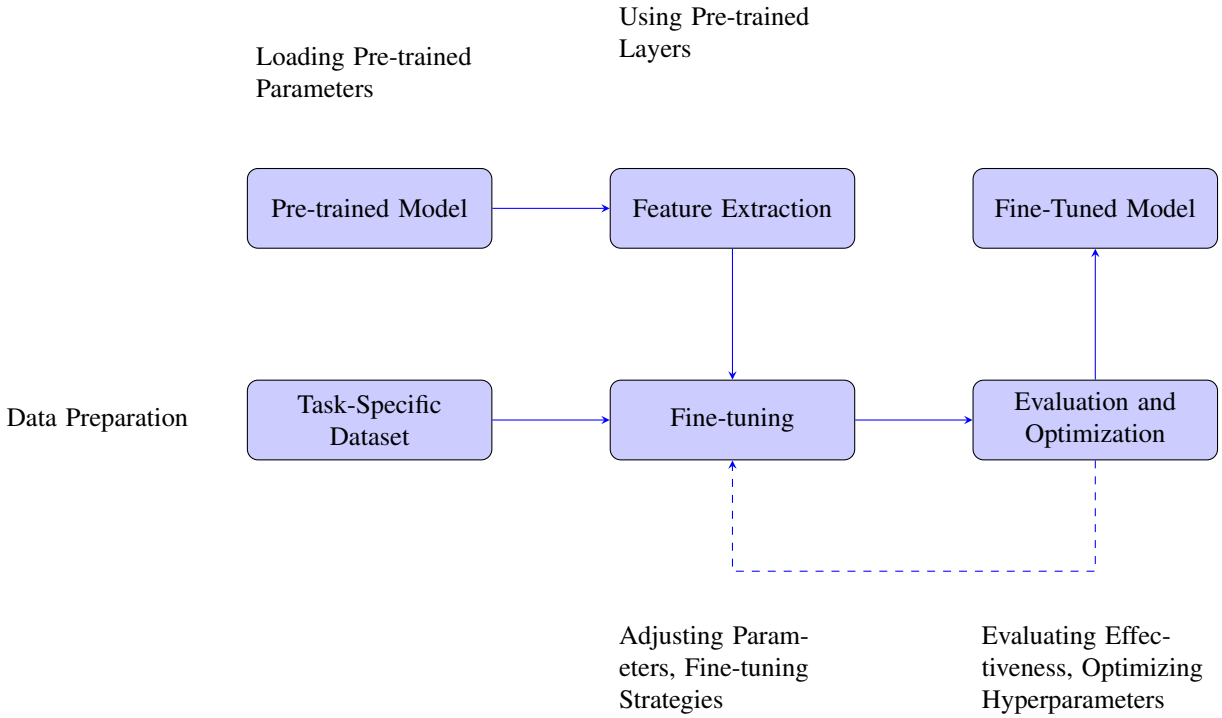

Fig. 1. Fine-tuning Process Schematic.

on Ukrainian texts, was chosen due to the small dataset size and the predominance of Ukrainian language.

The selected models were fine-tuned on the prepared text corpora using the `transformers` library from Hugging Face. The developed Python code allows loading pre-trained models, fine-tuning them on user-provided text data, and saving the fine-tuned models for further use. The total training time for the final experiment was 37 hours (average duration of one training iteration was 43 minutes) using a GPU on a local computer with a GeForce RTX 3080 (10 GB) graphics card.

C. Model Testing and Evaluation

The performance of the fine-tuned models was tested for generating chatbot responses to thematically related user queries. A graphical user interface was developed using the `gradio` library for user interaction with the chatbot (Figure 2).

The generated responses were evaluated using both automatic metrics (BERTScore, perplexity, BLEU, ROUGE) and human assessment, considering aspects such as relevance, coherence, factual correctness, and overall quality. The evaluation results were compared for models fine-tuned on different datasets and with different architectures.

IV. Results and Discussion

A. Dataset Analysis

The analysis of the created datasets revealed significant differences in their characteristics (Table II). The "CEUR Workshop Proceedings" dataset is much larger in volume

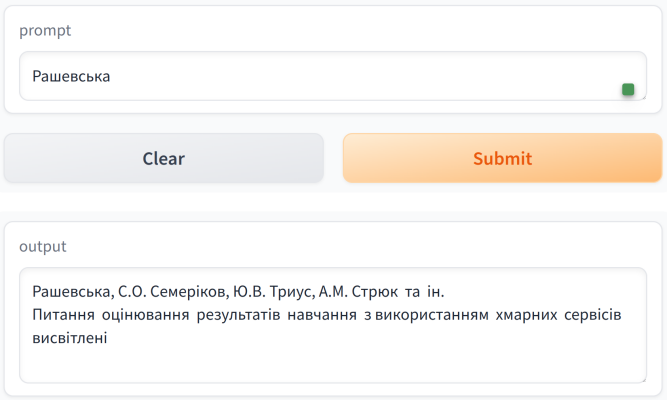

Fig. 2. Chatbot prototype interface for the GPT-2 model fine-tuned on the "Information Technologies and Learning Tools" dataset.

and covers a broader range of computer science topics, with English being the predominant language. The "Information Technologies and Learning Tools" dataset is smaller and more focused on the use of ICT in education, with Ukrainian being the main language.

These differences allowed for a comparative analysis of the effectiveness of chatbot models trained on heterogeneous text corpora.

B. Model Performance

The fine-tuned models demonstrated the ability to generate relevant and coherent responses to domain-specific user queries. The quality of the generated responses largely de-

979-8-3315-4263-4/24/$31.00 ©2024 IEEE

TABLE II
CHARACTERISTICS OF THE CREATED DATASETS

Characteristic	CEUR Workshop Proceedings	Information Technologies and Learning Tools
Volume	1917 MB	107 MB
Number of files	68,791	1,732
Time coverage	1995-2024	2006-2024
Thematic focus	Computer science and engineering	ICT in education
Main languages	English (94.8%), Russian (1.4%), German (1.1%)	Ukrainian (52.5%), Russian (26.7%), English (20.7%)

TABLE III
EVALUATION RESULTS FOR FINE-TUNED MODELS

Model	BERTScore	Perplexity	BLEU	ROUGE-L
GPT2-XL (CEUR)	0.85	15.3	0.32	0.41
gpt2-uk (ITLT)	0.79	22.7	0.26	0.35

pended on the volume and quality of the texts used for fine-tuning. The models fine-tuned on the larger "CEUR Workshop Proceedings" dataset generally produced more informative and contextually appropriate responses compared to the models trained on the smaller "Information Technologies and Learning Tools" dataset.

Table III presents the evaluation results for the fine-tuned models using automatic metrics. The GPT2-XL model fine-tuned on the "CEUR Workshop Proceedings" dataset achieved the highest scores across all metrics, indicating its superiority in generating relevant and coherent responses.

Human evaluation results were generally consistent with the automatic metrics, confirming the higher quality of responses generated by the GPT2-XL model fine-tuned on the "CEUR Workshop Proceedings" dataset. However, both models sometimes generated responses that were factually incorrect or inconsistent with the provided context, highlighting the need for further research on improving the robustness and reliability of chatbot models.

V. CONCLUSION

The conducted experiments revealed several challenges in applying transfer learning techniques to create specialised chatbots for Ukrainian scientific texts:

- Limited availability of large-scale, high-quality Ukrainian text corpora in specific domains.
- High computational costs associated with fine-tuning large language models.
- Difficulty in achieving factual correctness and consistency in generated responses.
- Need for more comprehensive evaluation metrics tailored to specialised domains.

Future research directions may include:

- Exploring advanced architectures, such as transformer variants optimised for low-resource languages and domain adaptation.
- Incorporating knowledge bases and external information sources to improve the factual accuracy of generated responses.
- Developing evaluation frameworks that consider the specific requirements and challenges of specialised domains.

ACKNOWLEDGMENT

The authors would like to thank the developers of the Hugging Face transformers library and the providers of the GPT-2 and gpt2-uk models for making their work publicly available. We also express our gratitude to the publishers of the "CEUR Workshop Proceedings" and "Information Technologies and Learning Tools" for maintaining open access to their valuable scientific content.

REFERENCES

[1] R. Liashenko and S. Semerikov, "The Determination and Visualisation of Key Concepts Related to the Training of Chatbots," in *Information Technology for Education, Science, and Technics: Proceedings of ITEST 2024, Volume 2*, ser. Lecture Notes on Data Engineering and Communications Technologies. Springer Cham, 2024, vol. 222. [Online]. Available: https://doi.org/10.1007/978-3-031-71804-5_8

[2] S. Patil, V. Mudaliar, and P. Kamat, "LSTM based Ensemble Network to enhance the learning of Long-term Dependencies in Chatbot," *International Journal of Automation and Smart Technology*, vol. 12, no. 1, pp. 2286–2286, Jan. 2022. [Online]. Available: https://doi.org/10.5875/ausmt.v12i1.2286

[3] Z. Ji, "A Multi-modal Seq2seq Chatbot Framework," in *Proceeding of 2021 International Conference on Wireless Communications, Networking and Applications*. Singapore: Springer Nature, 2022, pp. 225–233. [Online]. Available: https://doi.org/10.1007/978-981-19-2456-9_24

[4] D. Dharrao and S. Gite, "TherapyBot: a chatbot for mental well-being using transformers," *International Journal of Advances in Applied Sciences*, vol. 13, no. 1, pp. 1–12, Mar. 2024. [Online]. Available: http://dx.doi.org/10.11591/ijaas.v13.i1.pp1-12

[5] T.-L. Chou and Y.-L. Hsueh, "A Task-oriented Chatbot Based on LSTM and Reinforcement Learning," in *Proceedings of the 2019 3rd International Conference on Natural Language Processing and Information Retrieval*, ser. NLPIR '19. New York, NY, USA: Association for Computing Machinery, 2019, p. 87–91. [Online]. Available: https://doi.org/10.1145/3342827.3342844

[6] A. Kansal, "Finetuning: The Theory," in *Building Generative AI-Powered Apps: A Hands-on Guide for Developers*. Berkeley, CA: Apress, 2024, pp. 77–100. [Online]. Available: https://doi.org/10.1007/979-8-8688-0205-8_5

[7] Q.-D. L. Tran, A.-C. Le, and V.-N. Huynh, "Enhancing Conversational Model With Deep Reinforcement Learning and Adversarial Learning," *IEEE Access*, vol. 11, pp. 75 955–75 970, 2023.

[8] V. Ilievski, C. Musat, A. Hossman, and M. Baeriswyl, "Goal-Oriented Chatbot Dialog Management Bootstrapping with Transfer Learning," in *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*. International Joint Conferences on Artificial Intelligence Organization, 7 2018, pp. 4115–4121. [Online]. Available: https://doi.org/10.24963/ijcai.2018/572

[9] T. Taulli, *AI-Assisted Programming: Better Planning, Coding, Testing, and Deployment*. Sebastopol, CA: O'Reilly Media, Inc., 2024. [Online]. Available: https://www.oreilly.com/library/view/ai-assisted-programming/9781098164553/

[10] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language Models are Unsupervised Multitask Learners," Feb. 2019. [Online]. Available: https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf