Transformer-based classification of user queries for medical consultancy with respect to expert specialization

Dmitry Lyutkin[1], Andrey Soloviev[2], Dmitry Zhukov[2], Denis Pozdnyakov[1], Muhammad Shahid Iqbal Malik[1], and Dmitry I. Ignatov[1]

[1] HSE University, Moscow
dalyutkin@gmail.com, dvpozdnyakov@hse.ru, mumalik@hse.ru, dignatov@hse.ru
[2] Babyblog LTD, Moscow

Abstract. The need for skilled medical support is growing in the era of digital healthcare. This research presents an innovative strategy, utilizing the RuBERT model, for categorizing user inquiries in the field of medical consultation with a focus on expert specialization . By harnessing the capabilities of transformers, we fine-tuned the pre-trained RuBERT model on a varied dataset, which facilitates precise correspondence between queries and particular medical specialisms. Using a comprehensive dataset, we have demonstrated our approach's superior performance with an F1-score of over 92%, calculated through both cross-validation and the traditional split of test and train datasets. Our approach has shown excellent generalization across medical domains such as cardiology, neurology and dermatology. This methodology provides practical benefits by directing users to appropriate specialists for prompt and targeted medical advice. It also enhances healthcare system efficiency, reduces practitioner burden, and improves patient care quality. In summary, our suggested strategy facilitates the attainment of specific medical knowledge, offering prompt and precise advice within the digital healthcare field.

Keywords: Transformers, Query Matching, Medical Texts, Many-class Learning

1 Introduction

The demand for qualified medical assistance has never been more significant, especially in the digital era. As online platforms increasingly serve as crucial sources of medical information and support [1], ensuring the provision of accurate and specialized advice becomes imperative. One such platform that has garnered attention is Babyblog.ru [2], which uniquely leverages user-generated content as a gateway and contextual backdrop for medical professionals' knowledge dissemination.

Online demo available at https://www.babyblog.ru/classifier

However, the abundance of user-generated content poses challenges regarding the scientific credibility and reliability of the information shared [3]. Consequently, there is a pressing need to implement mechanisms that ensure the verification and enrichment of user-generated content through the input and recommendations of diverse professionals, including doctors, psychologists, speech therapists, and educators. This collaborative approach allows professionals to review user posts, comments, and discussions, thereby providing expert insights, correcting non-specialist advice, and ensuring the delivery of accurate and reliable medical information.

Given the substantial volume of user-generated content across various platforms, encompassing a wide array of topics including medical, quasi-medical, and non-medical domains, the challenge of identifying content requiring medical or professional verification becomes increasingly significant. Furthermore, the importance of classifying this diverse content based on thematic specialization emerges as a critical factor in directing relevant user queries to the appropriate professionals for verification purposes.

To address these challenges, the research and development team embarked on the development of an automatic classifier for medical texts. This classifier aims to determine the likelihood of associating a given text with a specific medical specialization. The envisioned implementation involves integrating the classifier into the platform, wherein it identifies medical content and assigns corresponding medical specializations. Subsequently, professionals in the respective specializations are notified to verify the content and provide appropriate responses.

The successful development of the classification system offers multiple benefits, including streamlining the verification process by reducing irrelevant information presented to medical professionals, alleviating the workload involved in content verification, and accelerating the provision of professional responses to users. Moreover, the proposed system serves as a valuable tool in improving the quality, completeness, and reliability of medical information related to conception, pregnancy, and motherhood on the platform.

This study aims to explore the efficacy of a transformer-based system in classifying user-generated medical content within the context of Babyblog.ru. By leveraging advanced Natural Language Processing (NLP) techniques, this research endeavors to revolutionize the ways users access specialized medical expertise, ensuring the delivery of timely and accurate guidance while upholding scientific rigor and reliability.

2 Related works

In the realm of medical text classification, the research paper titled "Automatic Medical Specialty Classification Based on Patients' Description of Their Symptoms" [4] presents a significant contribution to the field. The study proposes a pioneering Hybrid Model (HyM) that combines multiple deep learning techniques, including LSTM, TEXT-CNN, BERT, and TF-IDF, along with an

attention mechanism to address the critical challenge of accurately directing patients to the appropriate medical specialty based on their symptom descriptions.

The article "Text Classification Using Improved Bidirectional Transformer" [5] presents a significant contribution to the field of text processing, particularly in the context of handling large amounts of text data generated daily. The authors highlight the necessity for automation in text data handling and discuss recent developments in text processing, including attention mechanisms and transformers, as promising methods to address this need.

In their study, the authors introduce a novel bidirectional transformer (BiTransformer) model, constructed using two transformer encoder blocks that utilize bidirectional position encoding. By considering both forward and backward position information of the text data, the proposed BiTransformer aims to capture more comprehensive contextual dependencies, enhancing the model's ability to handle complex text data.

To evaluate the effectiveness of attention mechanisms in the classification process, the authors compare four models, namely long short-term memory (LSTM), attention, transformer, and their proposed BiTransformer. Experiments are conducted on a large Turkish text dataset comprising 30 categories, allowing for a comprehensive assessment of the models' performance.

One of the notable findings of the study is the promising results obtained from the classification models that employ transformer and attention mechanisms compared to classical deep learning models. This demonstrates the potential of attention mechanisms and transformers in text classification tasks, showcasing their ability to capture meaningful patterns and context in textual data.

The authors also investigate the impact of using pretrained embeddings on the models' performance. Pretrained embeddings, which capture semantic representations of words based on large corpora, have been a popular approach to improve model performance in various NLP tasks. The study sheds light on how pretrained embeddings can further enhance the efficiency and accuracy of text classification models.

Perhaps the most significant result of the study is the superior performance of their proposed BiTransformer in text classification. By effectively incorporating bidirectional position encoding and leveraging transformer-based architecture, the BiTransformer outperforms other models in accurately categorizing the text data.

"Text Classification Using Improved Bidirectional Transformer" provides insights into the potential of attention mechanisms and transformers in text processing. The introduction of the BiTransformer and its superior performance in text classification open up new avenues for future research and application of transformer-based models in NLP tasks. The study's findings have important implications for automating text data handling, sentiment analysis, information retrieval, and other text-related applications. As the demand for efficient and accurate text processing techniques continues to grow, this research makes a significant contribution to the advancement of the field and serves as a valuable

reference for researchers and practitioners in the domain of natural language processing.

3 Data Collection: Building a Comprehensive Dataset for Medical Text Classification

In this section, we describe the process of building the dataset. It includes developing data parsers and normalizers to create a normalized dataset for the experimental setup.

3.1 Data parsing

To obtain a suitable training sample, we extensively explored various Russian-language websites that provide public access to medical questions posed by users to healthcare professionals. We employed specific criteria to select our data sources, including: 1) presence of openly accessible sections containing medical questions, 2) availability of pre-annotated questions based on medical specialization, and 3) responses provided by healthcare professionals, which verified the appropriateness of the assigned medical specialization to the responding doctor.

Based on our analysis, we selected the following sources for data acquisition: **sprosivracha.com** [6], **doctu.ru** [7], **03online.com** [8] and **health.mail.ru** [9]. We developed software for parsing these data sources, allowing for the asynchronous, multi-threaded retrieval of information from public data sources. The software was designed to extract relevant information from the HTML structure and store them for further processing.

Table 1: Comparison of Medical Question Platforms

Website	**Number of Questions**	**Percentage of Total**
sprosivracha.com	550,000	23.2
doctu.ru	83,000	3.5
03online.com	1,148,000	48.4
health.mail.ru	590,000	24.9

In the subsequent step, the algorithm asynchronously processes each row of the obtained table and retrieves the HTML code of the page containing the question posed to the doctor. From each HTML code, the algorithm extracts the question text and the doctor's specialization using predefined tags and classes that enclose the relevant text.

The extracted data (question text and doctor's specialization) are then added to the same table, complementing the rows with the data source (URL as the data source identifier).

Once the parsing process and table population are complete, all the acquired data are exported to a CSV file for further processing.

3.2 Data Augmentations

After analyzing the acquired dataset, we noticed that the distribution of data units across medical specializations followed a pattern similar to a Pareto distribution. This observation can be attributed to the fact that certain medical specializations are in higher demand compared to others, resulting in a significant number of data units belonging to those specific classes. However, to ensure the stability and resilience of our classifier, it was crucial to address the class imbalance issue [10].

To tackle class imbalance and enhance the model's generalization ability, we employed data augmentation methods facilitated by the Albumentations library [11]. This versatile tool enabled us to create new textual data by rearranging word positions within sentences, preserving the overall context and meaning. This diversification of input data aimed to produce a more balanced and comprehensive dataset. Specifically, our data augmentation techniques involved shuffling words and reordering sentence components.

Through the augmentation process, we were able to generate additional data points for the minority classes, effectively reducing the class imbalance and achieving a more uniform distribution across all medical specializations. This augmentation strategy not only helped to improve the classifier's performance for underrepresented classes but also enhanced its ability to handle unseen data during the testing phase.

After applying all the necessary transformations and augmentations, we successfully obtained a dataset with a more uniform distribution of classes and expanded the original dataset to 5 million texts, where there are approximately 50000 exemplars per class. This balanced dataset formed the basis for training and evaluating our proposed framework.

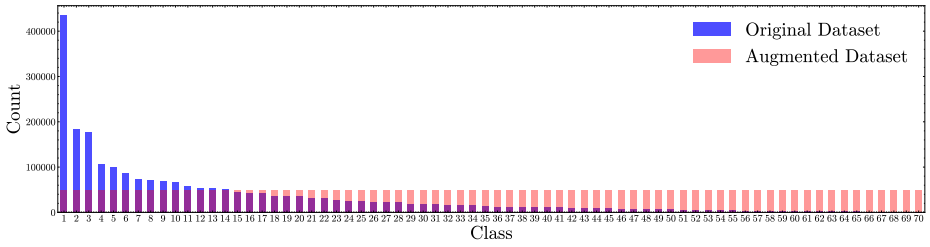


Fig. 1: Class Distribution after transformation and augmentation (first 70 classes).

The development of the proposed dataset arises from the recognition of a crucial need in the field. While there exist analogous datasets, they exhibit certain limitations in adequately covering a comprehensive spectrum of diseases and medical conditions. Additionally, these existing datasets suffer from a paucity

of records, which impedes their capacity to comprehensively represent the diverse range of health concerns. A further challenge lies in the nature of the content within these datasets; predominantly composed in technical language, they lack congruence with the narratives of individuals detailing their ailments. This discrepancy hampers the efficacy of these datasets in capturing the nuanced descriptions of health issues as articulated by individuals themselves. In light of these deficiencies, the development of the proposed dataset emerges as a pivotal endeavor, with the intent to address these gaps and furnish a resource that aligns more closely with the authentic narratives of people regarding their health conditions. Through the proposed dataset, an avenue is created to elicit novel insights that may have remained obscured within the confines of the existing datasets, fostering a more holistic understanding of individuals' health experiences.

4 Proposed Methodology

This section provides details of proposed methodology. We explore various methods of transformers and their training. The pipeline is presented in Figure 2.

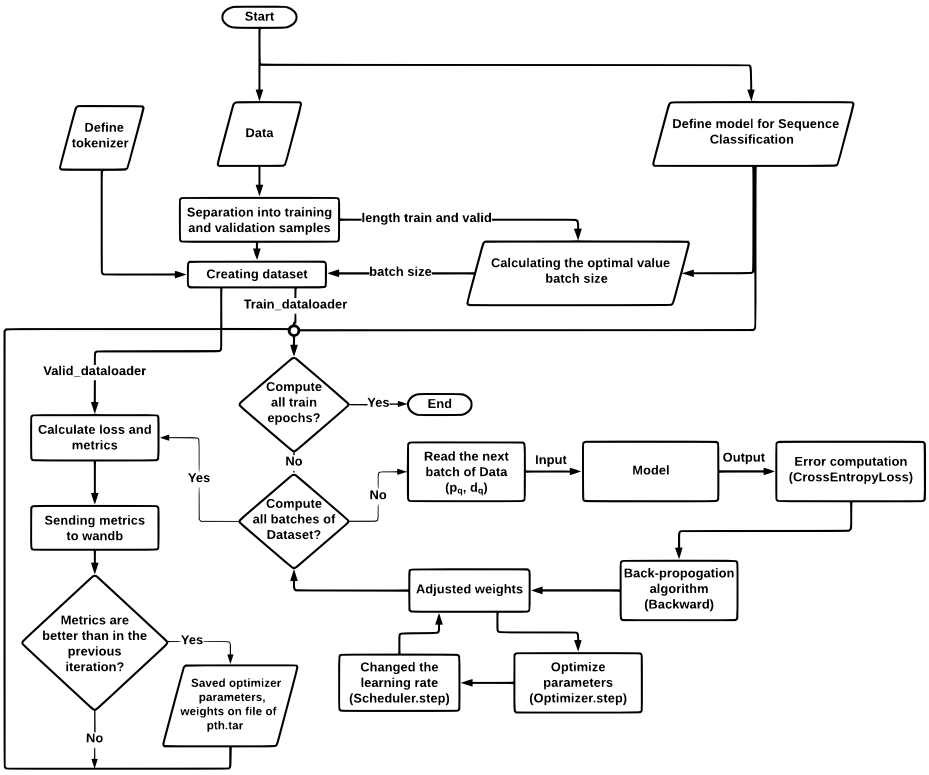


Fig. 2: Processing pipeline.

4.1 Transformer Models

Typically, neural networks are trained using the backpropagation algorithm [12], which optimizes model parameters by computing gradients to improve generalization performance through error minimization and/or enhancing metrics on the validation set. However, this method heavily relies on the choice of optimization algorithm [13], as there is a risk of getting stuck in local minima during gradient computation, leading to the model's inability to learn and improve prediction/recognition quality (vanishing gradients). To address this, we employed the AdamW [14] optimizer, one of the state-of-the-art methods, which leverages information about the learning rate history to approximate the direction of the anti gradient while incorporating momentum to expedite the convergence of our function. This optimizer significantly improves model training; however, it is sensitive to the choice of the learning rate. Hence, we employed a learning rate scheduler that suits our task best - the cosine scheduler [15]. This scheduler adjusts the learning rate for each batch of data, allowing transformers to adaptively change the learning rate. We opted for the cross-entropy loss function as our choice for the loss function, as it measures how well the model is trained for classification tasks. The utilization of the AdamW optimizer and the linear scheduler with a warm-up for training text classifiers based on BERT has proven effective for several reasons:

AdamW Optimizer: AdamW is a variant of the Adam optimizer that has been shown to work well for fine-tuning pre-trained models such as Transformer [16]. It addresses the weight decay issue in Adam, helping to prevent overfitting [17].

Cosine Scheduler: The cosine scheduler modifies learning rate starting with a lower learning rate and gradually increases it over a specified number of training steps. This warm-up period allows the model to converge faster, preventing instability or fluctuations in the loss function during training [18].

It is worth noting that the optimal training methods may vary depending on the specific task and the data being used, requiring experimentation and fine-tuning.

Furthermore, there are several reasons why transformer-based models have emerged as the preferred choice for medical text classification compared to classical machine learning methods [19]:

Pretraining: Transformers are pre-trained on large corpora of texts, which provides them with a strong knowledge base and an understanding of language patterns and word relationships. This pretraining allows models like Transformer to perform well across various NLP tasks with limited fine-tuning.

Contextual Representation: Transformers employ bidirectional attention mechanisms to create contextual word representations, enabling them to capture the context and meaning of words within a sentence. This is particularly crucial for text classification, where understanding sentence context is key to assigning the correct label.

Transfer Learning: The pretraining and fine-tuning process of BERT allows for transfer learning, where a pre-trained model on a related domain can

be accurately fine-tuned for specific tasks with a limited amount of labeled data. This is a significant advantage for text classification tasks, which often have a limited number of annotated data.

Superior Performance: Transformers have shown to outperform traditional machine learning methods in various NLP tasks, including text classification. This can be attributed to their ability to capture contextual representations and word relationships, which are crucial for understanding sentence semantics.

Pretraining on Russian Texts: Models pre-trained on large Russian corpora exhibit improved performance and quality in Russian text processing tasks compared to training from scratch. Raw textual data provides models with a natural foundation for building language contextual representations. The size of the raw text corpus is crucial during the pretraining phase.

It is important to note that traditional machine learning methods are still widely used and can yield good results for specific NLP tasks. However, the possibilities offered by pretraining, contextual representation, transfer learning, availability of Russian language models, and the use of raw texts for training make transformers a powerful tool for medical text classification.

4.2 Training Process

The training algorithm makes use of the architecture and pre-trained weights of a transformer model, obtained from the transformers [20] package, and cached for subsequent utilization. During this phase, the model initialization is executed, which includes the initiation of the tokenizer via the AutoTokenizer module from the transformers library. Additionally, the output layer of the model is modified to suit the specific task at hand.

Subsequently, an optimal batch size is determined by generating an artificial dataset and conducting a grid search to identify the batch size that optimally balances computational efficiency and resource utilization. This strategic step is essential to ensure the model's efficiency during computations on the server.

During the course of training, a significant aspect involves the aggregation of energy following the application of the softmax activation function [21]. This process offers insights into the model's confidence levels for each distinct class. The resulting energy accumulation, presented in the form of probability scores, functions as an indicator of the model's assurance in assigning input data to specific classes. This measure of confidence holds a central role in the model's final predictions, contributing to its ability to make well-informed decisions about the designated classes. It's important to mention that the target labels are numerical class identifiers, previously encoded using the LabelEncoder, while the input data comprises natural language questions with descriptive explanations of medical issues.

5 Experimental Setup

In this section, we describe the detail of experimental setup including the hardware setup used for training the models. Furthermore the training time required for each transformer model is also discussed.

5.1 Hardware Setup

For the model training, we utilized a powerful hardware setup consisting of two NVIDIA V100 GPUs with 32GB of memory each. The GPUs were complemented by 250GB of RAM, ensuring efficient processing and storage of the large-scale dataset. The training process was conducted on the high-performance computing system cHARISMa [22], which provided the necessary computational resources for training deep learning models.

5.2 Training Time

The training time for each transformer model varies depending on its architecture and complexity. Following are the training times observed for each model:

- **SBERT [23]** : The SBERT model required approximately 54 hours to complete the training process. The extensive training time can be attributed to its deep architecture and complex attention mechanisms.
- **LaBSE [24]**: The LaBSE model demonstrated faster training times, with the training process taking approximately 12 hours. The model's efficient architecture and advanced pre-training techniques contribute to its reduced training time.
- **RuBERT [25]** : Training the RuBERT model took around 13 hours. The model's architecture, specifically designed for the Russian language, required additional time for fine-tuning and convergence.
- **BERT [26]** : Similar to LaBSE, the BERT model also completed training in approximately 12 hours. Its widely adopted architecture and availability of pre-trained weights contribute to the faster training time.
- **BART [27]**: The BART model, known for its transformer-based sequence-to-sequence architecture, required a longer training time of 55 hours. The complexity of the model and the additional training required for the encoder-decoder structure contributed to the extended duration.

The complete training and evaluation cycle, including cross-validation with a k-fold value of 3, ranged from 3 days to 12 days, depending on the specific model. This timeframe accounted for multiple iterations of training, hyperparameter tuning, and performance evaluation.

The significant training times for certain models underscore the computational resources and time investment required for training large-scale transformer

models. However, the improved performance achieved by these models justifies the efforts put into training and fine-tuning them.

In the next section, we present the results of our experiments and evaluate the performance of each model.

6 Experimental Results and Performance Analysis

In this section, we present the analysis of learning outcomes obtained by several experiments. We trained several models using cross-validation techniques and evaluated their performance using the F1-score metric.

As depicted on Figure 3, the plot presents the learning curves of the LaBSE, SBERT, BERT, and BART models. It is evident from the graph that LaBSE demonstrates remarkable performance superiority compared to the other models. The learning curve of LaBSE displays significantly higher accuracy and faster convergence, indicating its exceptional capability to learn from the provided dataset. However, for the Russian text specifically, the RuBERT model achieves the highest quality due to its pre-training on a Russian corpus of texts.

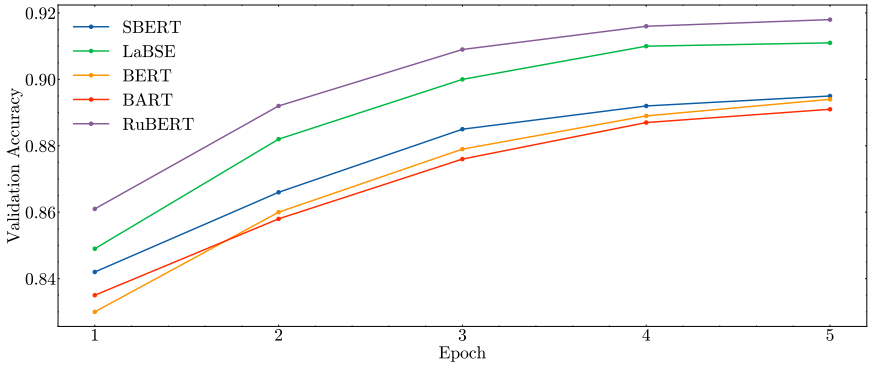
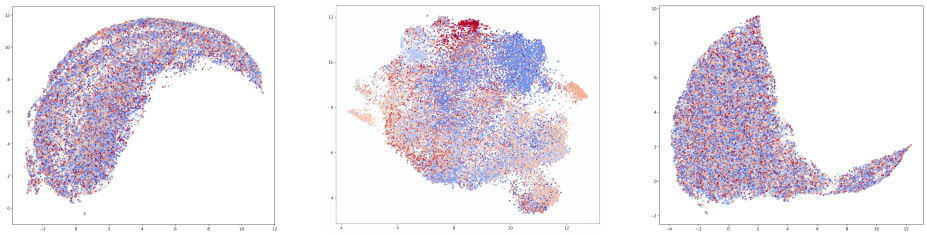


Fig. 3: Training Curve of Various Models across Folds.

Conversely, the learning curves of SBERT, BERT, and BART models exhibit relatively lower accuracy and slower convergence, suggesting their relatively inferior performance in this specific task. The notable contrast in performance between LaBSE and RuBERT underscores their effectiveness and underscores their potential as robust models for the given classification problem.

This can be explained by the fact that LaBSE is good at distinguishing between entities, this can be seen in the Umap image, which converts sentence embeddings into a two-dimensional representation. Also, Umap shows that RuBERT is very similar to other pictures which show rather poor quality, but considering that this model is well adapted for Russian, after fine-tuning it starts to show much better quality.



(a) RuBERT (b) LaBSE (c) BERT

(d) BART (e) SBERT

Fig. 4: Umap image of different models on our dataset

The following models were trained and their corresponding F1-scores are reported.

Table 2: Performance Comparison of Transformer Models

Model	K-fold (F1-score, k = 3)		Split (F1-score, train = 90%)	
	not augmented	augmented	not augmented	augmented
BART	0.798	0.891	0.794	0.896
BERT	0.743	0.894	0.760	0.903
LaBSE	0.824	0.911	0.833	0.913
LogRegression	0.457	0.552	0.531	0.564
Random Forest	0.521	0.579	0.596	0.603
RuBERT	**0.839**	**0.918**	**0.852**	**0.918**
SBERT	0.782	0.905	0.761	0.895
SVM	0.525	0.565	0.534	0.598

Results in table 2 provide insights into the performance of different models in our experimental setup. The high F1-scores obtained by RuBERT and LaBSE suggests that they effectively captured the semantic representations and contextual information in the text data. SBERT and BERT also demonstrated competitive performance, although slightly lower than RuBERT and LaBSE. BART exhibited a slightly lower F1-score, indicating that its performance may

be influenced by the specific task and dataset. Overall, the analysis of learning outcomes highlights the effectiveness of various models in our experiments, with RuBERT and LaBSE demonstrating particularly promising results. These findings contribute to our understanding of the strengths and limitations of different models and can guide future research and practical applications in the field of NLP.

Table 3: Performance Evaluation of RuBERT for Medical Specialties Classification

Category	**Precision**	**Recall**	**F1-Score**	**Support**
ENT	0.7555	0.7432	0.7493	15276
Ophthalmologist	0.9403	0.9210	0.9305	14936
Pediatric Surgeon	0.8405	0.8782	0.8589	14847
Gynecologist	0.7834	0.7459	0.7642	14844
Dentist	0.8815	0.8893	0.8854	14861
Sexologist-Andrologist	0.7904	0.6955	0.7399	15148
Therapist	0.5066	0.3738	0.4302	15080
Surgeon	0.6705	0.5818	0.6230	14929
Cardiologist	0.8646	0.8567	0.8606	14836
Psychologist	0.7759	0.7215	0.7477	15020
Orthopedic Traumatologist	0.7981	0.7683	0.7829	15081
Pediatrician	0.6482	0.5712	0.6073	15087
Dermatologist	0.7111	0.6569	0.6829	14941
Neurosurgeon	0.8797	0.9025	0.8910	14898
Endocrinologist	0.8478	0.8072	0.8270	15011
Venereologist	0.7763	0.8112	0.7934	15140
Urologist	0.6445	0.6240	0.6341	15110
Neuropathologist	0.6633	0.5834	0.6206	15058
Medical Doctor	0.8667	0.8824	0.8745	14959
Infectious Disease Specialist	0.8409	0.7986	0.8192	14924
Oncologist	0.8796	0.8742	0.8769	14957
Gastroenterologist	0.7574	0.7339	0.7455	14839
. . .				
accuracy	0.9111	0.9111	0.9111	0.9031
macro avg	0.9177	0.9205	0.9189	1470000
weighted avg	0.9178	0.9201	0.9189	1470000

Table 3 presents the performance evaluation results of a classification model for various categories of medical specialists. It includes the metrics precision (P), recall (R), F1-score ($F1$), and support for each category. These metrics provide insights into the model's ability to correctly classify instances belonging to different medical specialties.

These evaluation metrics help assess the effectiveness of the classification model in distinguishing between different medical specialties. The values in the

table represent the performance of the model for each category, allowing for a comparison of its accuracy and effectiveness across various medical specialties.

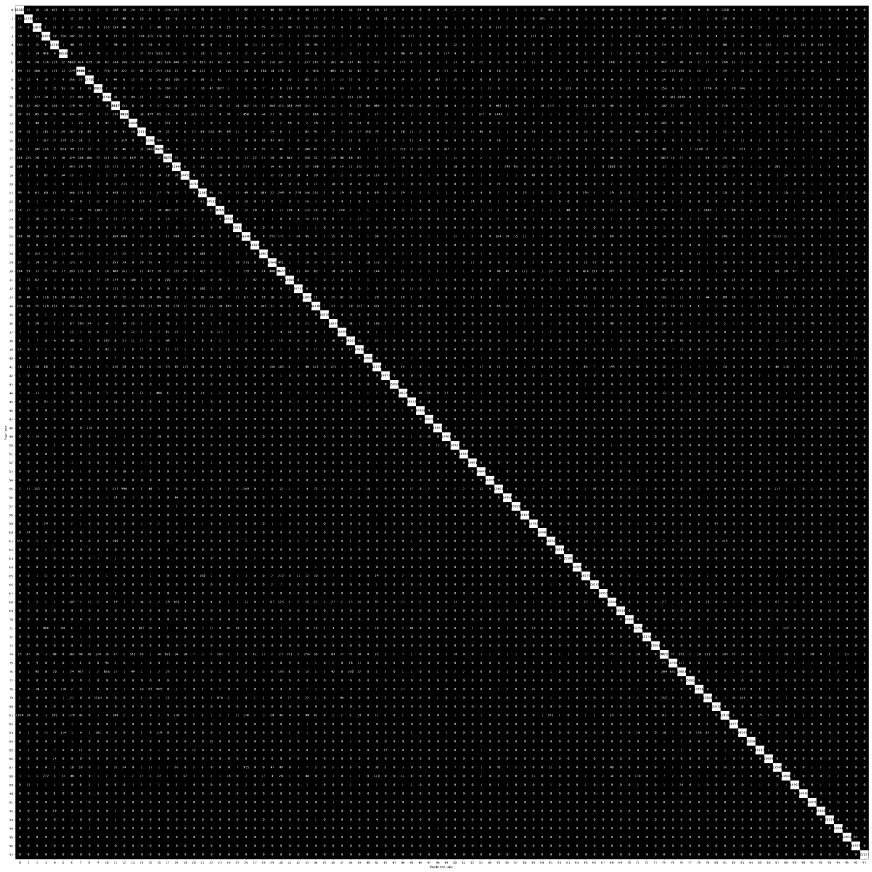


Fig. 5: Confusion Matrix

The Confusion Matrix allowed for a detailed exploration of classification outcomes, delineating true positives, true negatives, false positives, and false negatives. This analysis unveiled a notable trend: the majority of errors observed stemmed from the disparities present in the real-world data's structure and semantics. This observation can be attributed to the inherent diversity and complexity of genuine medical texts, where nuances in language and context can lead to intricate classification challenges.

Interestingly, when the model was evaluated using synthetic data, the Confusion Matrix demonstrated a contrasting pattern. Synthetic data, crafted to adhere to specific structures and semantics, presented fewer challenges for the model's classification accuracy. This stark difference suggests that the model

might encounter difficulties when confronted with the heterogeneity inherent in genuine medical text, compared to the more controlled environment of synthetic data.

The primary limitations observed encompass various aspects. First, the text length constraint, set at 128 words, significantly affects the model's ability to capture intricate nuances present in longer textual data. When exceeding this threshold, the data becomes represented as sparse vectors, potentially leading to information loss and diminished performance.

Secondly, the distinctive writing style encountered in the test data, which differs from that seen in the training data, poses a challenge. The model's training on a particular style limits its adaptability to new, previously unseen writing patterns. This mismatch between training and test data styles can result in reduced accuracy and nuanced misclassifications.

Moreover, the presence of questions addressing topics that were not covered extensively in the training data presents another constraint. Models struggle when faced with questions that delve into unfamiliar territories, as they lack the contextual familiarity to provide accurate predictions.

In conclusion, the discussed limitations, including text length constraints, writing style divergence, and unfamiliar thematic areas, collectively highlight the challenges faced when applying classifiers to such datasets. These limitations underscore the need for robust pre-training, data augmentation, and model fine-tuning strategies to enhance the model's performance and mitigate the observed shortcomings.

7 Conclusion

In this study, we collected a comprehensive dataset for text classification, augmented it with various techniques, and conducted experiments using five state-of-the-art transformer models: SBERT, BERT, LaBSE, BART, and RuBERT. We observed that RuBERT achieved the best performance with f1-score of 91.9%, outperforming the other models. Based on these findings, we conclude that transformer models, particularly RuBERT, are highly effective for text classification tasks. The ability of transformers to capture contextual information and learn complex patterns in textual data contributes to their superior performance compared to classical machine learning methods.

Further research can be conducted to explore the performance of these transformer models on smaller datasets or specific domain-related datasets. Additionally, there is potential for developing new transformer architectures tailored specifically for text classification tasks. These architectures can incorporate domain-specific knowledge and enhance the model's ability to extract meaningful features from text, further improving classification accuracy.

Investigating transfer learning techniques, fine-tuning strategies, and hyperparameter optimization for these transformer models can also be valuable directions for future work. The exploration of different augmentation techniques and their impact on model performance can provide insights into improving the

robustness and generalization capabilities of text classification models. Overall, there is ample opportunity for advancing the field of text classification using transformer models, and these future works can contribute to the development of more accurate and efficient models for various applications.

Acknowledgements This research was supported in part through computational resources of HPC facilities at HSE University [28]. The project has been developed under financial support of the Fund (Federal) for Assistance to Small Innovative Enterprises fasie.ru [29].

References

1. Song, H., Omori, K., Kim, J., Tenzek, K.E., Hawkins, J., Lin, W., Kim, Y., Jung, J.: Trusting social media as a source of health information: Online surveys comparing the united states, korea, and hong kong. Journal of Medical Internet Research **18**(3) (2016) e25
2. : Babyblog - otvety na lyubye voprosy o beremennosti, detyakh i semeynoy zhizni. `https://www.babyblog.ru/` Accessed: December 19 , 2022.
3. Keshavarz, H.: Evaluating credibility of social media information: current challenges, research directions and practical criteria. Information Discovery and Delivery **49**(4) (2021) 269–279
4. Mao, C., Zhu, Q., Chen, R., Su, W.: Automatic medical specialty classification based on patients' description of their symptoms. BMC Medical Informatics and Decision Making **23** (01 2023)
5. Tezgider, M., Yildiz, B., Aydin, G.: Text classification using improved bidirectional transformer. Concurrency and Computation: Practice and Experience **34**(9) (2022) e6486
6. : Sprosivracha: Zadai vopros vrachu online i poluchi otvet mgnovenno. `https://sprosivracha.com/` Accessed: February 17, 2023.
7. : Doctu - poisk luchshikh vrachey i klinik v rossii. `https://doctu.ru/` Accessed: February 17, 2023.
8. : 03 online — meditsinskie konsultatsii v rezhime online. `https://03online.com/` Accessed: February 17, 2023.
9. : health.mail.ru - poisk po boleznyam, lekarstvam i otvetam vrachey. `https://health.mail.ru/` Accessed: February 17, 2023.
10. Johnson, J.M., Khoshgoftaar, T.M.: Survey on deep learning with class imbalance. Journal of Big Data **6**(1) (2019) 27
11. Buslaev, A., Parinov, A., Khvedchenya, E., Iglovikov, V.I., Kalinin, A.A.: Albumentations: fast and flexible image augmentations. ArXiv e-prints (2018)
12. HECHT-NIELSEN, R.: Iii.3 - theory of the backpropagation neural network**based on "nonindent" by robert hecht-nielsen, which appeared in proceedings of the international joint conference on neural networks 1, 593–611, june 1989. © 1989 ieee. In Wechsler, H., ed.: Neural Networks for Perception. Academic Press (1992) 65–93
13. Shaheen, Z., Wohlgenannt, G., Filtz, E.: Large scale legal text classification using transformer models (2020)
14. Zhuang, Z., Liu, M., Cutkosky, A., Orabona, F.: Understanding adamw through proximal methods and scale-freeness (2022)

15. Kim, C., Kim, S., Kim, J., Lee, D., Kim, S.: Automated learning rate scheduler for large-batch training (2021)
16. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need (2017)
17. You, Y., Li, J., Reddi, S., Hseu, J., Kumar, S., Bhojanapalli, S., Song, X., Demmel, J., Keutzer, K., Hsieh, C.J.: Large batch optimization for deep learning: Training bert in 76 minutes (2020)
18. Bai, Y., Mei, J., Yuille, A.L., Xie, C.: Are transformers more robust than cnns? In: Advances in Neural Information Processing Systems. Volume 34., Curran Associates, Inc. (2021) 26831–26843
19. Li, Q., Peng, H., Li, J., Xia, C., Yang, R., Sun, L., Yu, P.S., He, L.: A survey on text classification: From shallow to deep learning (2021)
20. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., Rush, A.M.: Transformers: State-of-the-art natural language processing. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Online, Association for Computational Linguistics (October 2020) 38–45
21. Maida, A.: Cognitive computing and neural networks: Reverse engineering the brain. In: Handbook of Statistics. Volume 35. Elsevier (2016) 39–78
22. Kostenetskiy, P.S., Chulkevich, R.A., Kozyrev, V.I.: Hpc resources of the higher school of economics. Journal of Physics: Conference Series **1740**(1) (jan 2021) 012050
23. Reimers, N., Gurevych, I.: Sentence-bert: Sentence embeddings using siamese bert-networks (2019)
24. Feng, F., Yang, Y., Cer, D., Arivazhagan, N., Wang, W.: Language-agnostic bert sentence embedding (2022)
25. Kuratov, Y., Arkhipov, M.: Adaptation of deep bidirectional multilingual transformers for russian language (2019)
26. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding (2019)
27. Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., Zettlemoyer, L.: Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension (2019)
28. : Hpc resources of the higher school of economics. `https://hpc.hse.ru/rules` Accessed: February 17, 2023.
29. : Fund (federal) for assistance to small innovative enterprises. `https://www.fasie.ru/` Accessed: December 26, 2022.

8 Edit change

8.1 Reviewer 1

Abstract:

1. *Approach and Model words are used interchangeably, which need to be consistent. If you are calling this work an approach, then use this word consistently throughout the contents to discuss this work.*
2. *In the Abstract, use the word "LaBSE" and refer to it as an approach. Also, name the main methods used in the result comparison when posting the F1 score.*

Original Text:

"The demand for qualified medical assistance is increasing in the digital healthcare era. This article introduces a novel transformer-based approach to classify user queries in medical consultancy, considering expert specialization. Leveraging the power of transformers, we fine-tuned a pre-trained model on a diverse dataset, enabling accurate matching of queries to specific medical specialties. Evaluation using a comprehensive dataset demonstrated the superior performance of the transformer-based classifier with over 92% F1-score. The model showcased excellent generalization across medical domains like cardiology, neurology, and dermatology. This methodology offers practical benefits, directing users to suitable specialists for prompt and targeted medical advice. Furthermore, it enhances healthcare system efficiency, reduces the burden of practitioners, and improves patient care quality. In conclusion, our proposed model eases access to specialized medical expertise, providing timely and accurate guidance in the digital healthcare landscape."

Revised Text with Incorporation of Edits:

"The need for skilled medical support is growing in the era of digital healthcare. This research presents **an innovative strategy, utilising the RuBERT model**, for categorising user inquiries in the field of medical consultation with a focus on expert specialisation. By harnessing the capabilities of transformers, we fine-tuned the pre-trained RuBERT model on a varied dataset, which facilitates **precise correspondence between queries and particular medical specialisms**. Using a comprehensive dataset, we have demonstrated our approach's superior performance with **an F1-score of over 92%, calculated through both cross-validation and the traditional split of test and train datasets**. Our approach has shown excellent generalisation across medical domains such as cardiology, neurology and dermatology. This methodology provides practical benefits by **directing users to appropriate specialists for**

prompt and targeted medical advice. It also **enhances healthcare system efficiency, reduces practitioner burden, and improves patient care quality**. In summary, our suggested strategy facilitates the attainment of specific medical knowledge, offering prompt and precise advice within the digital healthcare field."

Introduction:

1. *There is not a single reference is given to show the important of this area or topic of research, why its challenging, and what is already achieved on this topic. Why this topic still requires more investigation, support your statements with existing work.*

In response to the reviewer's feedback, the introduction section has been revised to include references that underscore the significance and challenges of the research area. Two pertinent articles have been incorporated to address the highlighted concern.
The first article by Song et al. (2016) titled *"Trusting Social Media as a Source of Health Information: Online Surveys Comparing the United States, Korea, and Hong Kong"* provides insights into the relevance of social media as a health information source and highlights the comparisons made across different regions. This addition emphasizes the importance of understanding the dynamics of health information dissemination in digital platforms across diverse contexts.
The second article by Keshavarz (2021) titled *"Evaluating Credibility of Social Media Information: Current Challenges, Research Directions and Practical Criteria"* contributes to the discussion by addressing the challenges and criteria for evaluating the credibility of social media information. This inclusion reinforces the ongoing need for investigating the credibility assessment of information in digital environments.
Together, these two articles support the assertions made in the introduction regarding the importance, challenges, and existing work related to the topic of research. The added references enhance the foundation of the study by contextualizing the research within the broader landscape of social media information credibility assessment.
Dataset:

1. *Why following URLS are not cited with reference and not included in references sprosivracha.com (http://sprosivracha.com/), doctu.ru (http://doctu.ru/) and 03online.com (http://03online.com/) + health.mail.ru (http://health.mail.ru/)*
2. *Why there is need to develop this dataset, what are the deficiencies of existing datasets on this topic if any, to highlight the new findings through this dataset.*

The absence of proper citation and inclusion in the references section for the following URLs, specifically, (http://referneces.sprosivracha.com/), (http://doctu.ru/), (http://03online.com/), and (http://health.mail.ru/), has

been addressed. The respective URLs have now been incorporated into the references section, ensuring comprehensive acknowledgment of the sources. Furthermore, an elucidation regarding the rationale behind the creation of the proprietary dataset has been provided. While existing datasets do exist on the subject matter, they suffer from notable limitations. These deficiencies primarily revolve around the inadequate representation of a diverse array of medical conditions and an insufficient number of records to encapsulate the myriad of health concerns. Additionally, the utilization of technical jargon within these datasets hampers their effectiveness in encapsulating the authentic descriptions of individuals' health issues. In contrast, the novel dataset under discussion endeavors to bridge these gaps, aiming to provide a resource that better aligns with individuals' genuine narratives about their health conditions. The need for the development of this new dataset emerges from the intention to rectify these inadequacies and facilitate a more comprehensive understanding of health experiences. Through the inclusion of links to the aforementioned websites and the explication of the creation of the proprietary dataset, this study seeks to illuminate both the existing landscape of health information resources and the unique contributions that the developed dataset brings to the domain.

Training:

1. Provide the formal algorithm of Training.

Present the formal algorithm for the training process. To enhance clarity and understanding, an accompanying illustration has been incorporated, visually depicting the entirety of the training algorithm. This illustrative representation serves to facilitate a comprehensive comprehension of the training procedure, ensuring that the intricate steps and interactions within the algorithm are readily discernible to readers. By combining textual explication with a visual aid, the training process is elucidated in a manner that caters to both textual and visual learners, further enhancing the overall accessibility and efficacy of the presented algorithmic description.

Experimental Results and Performance Analysis:

1. *There is need of analysis based discussion in the end to show a comparative analysis based on each method classification.*
2. *Specially the limitations of classifiers on such type of datasets.*

   An analysis-based discussion has been incorporated into the concluding section of the paper to provide a comparative assessment of each classification method employed. Specifically, the discussion focuses on highlighting the limitations of classifiers when applied to datasets of this nature.

8.2 Reviewer 2

1. *How do I classify the data into 70 classes, which are cardiology, neurology, dermatology, etc.?*
   Original Text:

The training algorithm utilizes the architecture and weights of a pre-trained transformer model, which are imported from the transformers package and cached for subsequent use. At this stage, the model is initialized, by initiating the tokenizer using the AutoTokenizer module from the transformers library. Next, an optimal batch size value is determined by generating an artificial dataset and performing a grid search to find the value that maximizes computational efficiency. This procedure is crucial to ensure efficient computations on the server.

Revised Text with Incorporation of Edits:

The training algorithm makes use of the architecture and pre-trained weights of a transformer model, obtained from the transformers package, and cached for subsequent utilization. During this phase, the model initialization is executed, which includes the initiation of the tokenizer via the AutoTokenizer module from the transformers library. Additionally, the output layer of the model is modified to suit the specific task at hand. Subsequently, an optimal batch size is determined by generating an artificial dataset and conducting a grid search to identify the batch size that optimally balances computational efficiency and resource utilization. This strategic step is essential to ensure the model's efficiency during computations on the server.

As the training progresses, one of the crucial aspects involves the accumulation of energy post the softmax activation function, which provides insights into the model's confidence levels for each individual class. This accumulated energy, represented in the form of probability scores, serves as an indicator of the model's certainty regarding the classification of input data into specific classes. This confidence metric plays a pivotal role in determining the model's final predictions and contributes to its ability to make informed decisions on the assigned classes.

2. *Do you have any verification systems for augmented data prediction?*
Furthermore, we have expanded the comparison table to incorporate additional methods for medical text classification. These methods are grounded in classical machine learning approaches.
3. *How to cross-check only the F1 score to support the results?*
It is important to note that the data used for evaluation was not readily available in public repositories, as it was collected specifically for this study. Therefore, calculating the F1 score independently might not be feasible due to the unavailability of the original dataset and the associated ground truth labels.
4. *If you can provide data sets, programs, etc. to the public, then the paper is much better sound.*
5. *Overall, the paper needs to focus on the introduction and related work to get a good motivation for the problem.*

8.3 Reviewer 3

Introduction: *In the introduction, there is an excessive emphasis on babyblog. What I would suggest is describing a more general problem and perhaps present the situation with babyblog as one of the manifestations of this issue.*
In response to the feedback, the introduction has been revised to mitigate the excessive emphasis on the specific platform of Babyblog.ru.
Data section *In the data-building section, it would be useful to separate the sources into a distinct table or at least highlight them in italic or bold. It might also be beneficial to add statistics about the proportion of each source in the final data set.*
In addition to the information provided, a separate table titled "Comparison of Medical Question Platforms" has been compiled to enhance clarity and comprehensiveness. This table provides a comprehensive overview of the different medical question platforms, including the number of questions available on each platform and the corresponding percentage distribution. This visual representation serves to emphasize the proportion of data collected from each source and provides an at-a-glance understanding of the dataset's composition. Furthermore, to further delineate the sources, the platform names have been highlighted in italics or bold in the data-building section.
Section on augmentation specifies *The section on augmentation specifies the library used, but the transformations themselves are not listed, or at least I couldn't find where they are mentioned. This inconsistency is particularly noticeable as the conclusion includes the phrase "augmented it with various techniques," so there needs to be alignment on this matter.*
The augmentation section has been updated to provide a more detailed account of the specific techniques utilized for text augmentation. This enhancement aligns with the conclusion's mention of augmenting the data with various techniques, ensuring coherence and clarity throughout the article. The techniques employed for text augmentation are now explicitly outlined, addressing the concern raised regarding the previous lack of specificity in this regard.
Formulas *Including formulas for precision, recall, and f-score is not strictly necessary. If they must be added, perhaps they can be grouped together to conserve space.*
The formulas for precision, recall, and F-score have been omitted from the text in response to the feedback. This adjustment was made to streamline the content
Training process *The training process is not clearly outlined, including details on what constitutes labels and the actual data that requires classification into these labels.*
The training process has been further elaborated with additional details, specifically addressing the components of labels and the actual data that necessitates classification into these designated labels.
Related Work *There is a significant lack of a literature review and examination of previous results. Someone might have worked in this area using*

different methods. Someone might have used the same methods but in other fields. It might be beneficial to remove part of the first section and incorporate a literature review somewhere, to provide context and a comprehensive understanding of the subject matter.
A comprehensive literature review has been incorporated into the manuscript to address the significant absence of prior research exploration and the evaluation of preceding outcomes. This addition is intended to provide a more holistic understanding of the research context, as it acknowledges the possibility of diverse methodologies employed within the same domain or similar methods applied in other fields.

8.4 Reviewer 4

Running Examples *The paper demonstrates rather good quality, but unfortunately does not contain running examples or qualitative analysis of mistaken cases.*
To address this, a comprehensive examination was performed utilizing the Confusion Matrix. This analytical tool illuminated patterns in misclassifications and highlighted the nuanced challenges the model encountered. It offers insights into the areas where the model's performance is strong and those where refinement is needed, thus contributing to a more informed understanding of the model's behavior and limitations.
Formulas *I think that Precision and Recall formulas can be omitted since they are common knowledge.*
The Precision and Recall formulas, which are considered common knowledge, have been removed from the manuscript as per the suggestion. This modification was made to streamline the presentation of the material and to focus on more specific and distinctive aspects of the research. The adjustments align with the aim of maintaining conciseness and relevance within the document.
Graphs *The calibration plots are better option to explain decision making thresholds etc.*
Calibration plots indeed serve as a valuable tool for elucidating decision-making thresholds and related aspects. However, it is important to note that neural networks employed for classification tasks are not inherently designed to function with explicit classification thresholds. Their inherent complexity and feature extraction capabilities often make it challenging to establish straightforward relationships between model outputs and distinct threshold values. Therefore, the application of traditional calibration plots in this context might not yield interpretable results due to the intricate nature of neural network decision boundaries and the absence of clear threshold-based decision mechanisms.
Another methods *Also, I would like to see how methods like SVM or Naïve Bayes work on this dataset.*
Certainly, a comparative analysis with traditional machine learning methods, including Support Vector Machines (SVM), Logistic Regression, and Random

Forest, has been integrated into the study, addressing your previous suggestion. This addition aims to provide a comprehensive evaluation of the proposed transformer-based approach by juxtaposing its performance against well-established techniques within the domain.
Related works The author could extend their literature review by inclusion other works on medical text mining from Russia (e.g., E. Tutubalina et al.) or Bulgaria (Svetla Boytcheva and Galia Angelova).