Towards Automatic Medical Report Classification in Czech

Pavel Přibáň[a], Josef Baloun[b], Jiří Martínek[c], Ladislav Lenc[d], Martin Prantl[e]
and Pavel Král[f]

Department of Computer Science and Engineering, Faculty of Applied Sciences,
University of West Bohemia, Pilsen, Czech Republic

Keywords: Machine Learning, Classification, Multi-Label, Single-Label, Medical Data.

Abstract: This paper deals with the automatic classification of medical reports in the form of unstructured texts in Czech. The outcomes of this work are intended to be integrated into a coding assistant, a system that will help the clinical coders with the manual coding of the diagnoses.

To solve this task, we compare several approaches based on deep neural networks. We compare the models in two different scenarios to show their advantages and drawbacks. The results demonstrate that hierarchical GRU with attention outperforms all other models in both cases.

The experiments further show that the system can significantly reduce the workload of the operators and thus also saves time and money. To the best of our knowledge, this is the first attempt at automatic medical report classification in the Czech language.

1 INTRODUCTION

International classification of diseases (ICD) is a standard that assigns codes to diseases and other causes of patient encounters with the health care system. One of the main usages of the coding system is for billing data that hospitals report to insurance companies.

The coding thus should be done by well-trained and experienced staff. However, in the real world, this task is often performed by a doctor who writes the report, which can lead to inconsistencies and mistakes in the coding.

In the last several years, there have been efforts to solve this task automatically because manual coding is an expensive and time-consuming task, often erroneous due to the human factor. At least partial automation of this process will bring better reliability and especially time and money savings. The prediction models can also be used to validate already reported diagnoses.

Czech doctors typically write medical reports in the form of unstructured text. The texts are usually assigned with one main and several secondary diagnoses.

The main goal of this paper is to propose and compare different approaches to solve automatic diagnosis coding in Czech using deep neural networks. We compare and evaluate five deep models in two different scenarios: 1) main diagnosis classification; 2) all diagnoses classification. The approaches are compared with a simple baseline based on a multi-layer perceptron (MLP) to show their advantages and inconveniences.

For evaluation, we use a novel Czech medical dataset collected from a large Czech hospital. The dataset contains more than 300,000 anonymised medical reports associated with the ICD codes.

To the best of our knowledge, this work represents the first attempts at automatic medical report classification in the Czech language.

2 RELATED WORK

In this section, we summarise approaches that are used for multi-label classification in the medical domain as well as methods used in similar text categorisation tasks.

An approach for classifying legislative documents from the European Union was presented by

[a] https://orcid.org/0000-0002-8744-8726
[b] https://orcid.org/0000-0003-1923-5355
[c] https://orcid.org/0000-0003-2981-1723
[d] https://orcid.org/0000-0002-1066-7269
[e] https://orcid.org/0000-0002-7900-5028
[f] https://orcid.org/0000-0002-3096-675X

Přibáň, P., Baloun, J., Martínek, J., Lenc, L., Prantl, M. and Král, P.
Towards Automatic Medical Report Classification in Czech.
DOI: 10.5220/0011641900003393
In *Proceedings of the 15th International Conference on Agents and Artificial Intelligence (ICAART 2023) - Volume 3*, pages 228-233
ISBN: 978-989-758-623-1; ISSN: 2184-433X

(Chalkidis et al., 2019). (You et al., 2019) carried out the benchmark on the six most common multi-label datasets, including the huge Amazon-3M ((McAuley and Leskovec, 2013)) with their proposed deep model called AttentionXML.

In the medical domain, (Perotte et al., 2014) used models based on support vector machines (SVM) for predicting ICD codes from discharge summaries. Approaches based on recurrent neural networks for the same task were presented in (Shi et al., 2017; Vani et al., 2017).

(Mullenbach et al., 2018) proposed an attentional convolutional network CNN-LWAN for the prediction of medical codes and evaluated it on MIMIC-II and MIMIC-III ((Johnson et al., 2016)) datasets. The authors of (Baumel et al., 2018) investigated several models, including hierarchical attention GRU (HA-GRU), for predicting diagnosis codes. The best performance has been obtained by HA-GRU. Moreover, the sentence-level attention can be visualised to highlight important parts (words or sentences) of clinical documentation, which is beneficial for operators who check the outputs.

Authors of (El Boukkouri et al., 2020) state that Character level embeddings are better for medical data compared to word-level ones.

Diagnoses assigned to medical reports often correlate with each other. Some diagnoses appear together very often and some combinations are quite rare. This fact was addressed in (Xun et al., 2020) where the authors developed a special classification layer called CorNet which can be appended to arbitrary architecture.

Correlation is also used by Gu et al. (Gu et al., 2021). They use Graph Convolutional Network to find correlations between diagnoses. As a result, the system is also capable of predicting less frequent diagnoses with improved precision.

The above-mentioned approaches are evaluated mainly in English. However, to the best of our knowledge, no work for automatic diagnoses classification dealing with the Czech language exists.

3 APPROACHES

We propose and evaluate the following state-of-the-art models from the text classification domain. As a baseline, we use an MLP with an input based on the TF-IDF document representation.

For all models, numbers and diacritics are removed from the input text that is converted to lowercase.

3.1 Multi-Layer Perceptron

We use TF-IDF (term frequency-inverse document frequency) method for feature selection and document representation. The MLP has the following topology: 8000 nodes in the input layer, 8192 neurons in the hidden layer and the output layer dimension corresponds to the number of classes. This bag-of-word (BoW) model is hereafter called *MLP (base)*.

3.2 Convolutional Neural Network

We use the architecture proposed in (Lenc and Král, 2016), which is an adaptation of the CNN model presented by (Kim, 2014). The model contains an embedding layer with randomly initialised word embeddings which are tuned during the training process. This model is hereafter called *CNN 512* when we use 512 words for the input or *CNN 1024* if the input is composed of 1024 words.

3.3 ELECTRA

Another model for the classification that we selected is the Czech pre-trained *Small-E-Czech* model (Kocián et al., 2022). This is a Czech version of the English *ELECTRA-small* (Clark et al., 2020) model based on the Transformer (Vaswani et al., 2017) architecture. We decided to use this model since it has significantly fewer parameters (14M) than the other available Czech BERT-like models, Czert (Sido et al., 2021) (110M) or RobeCzech (Straka et al., 2021) (125M). Thus, it can be fine-tuned faster and with less GPU memory than the latter two. We fine-tune the model in the same way as the authors in the original ELECTRA (Clark et al., 2020) model, i.e., we add a classification head that consists of a simple linear classifier on top of ELECTRA.

3.4 Document Character-Level Embedding

The architecture utilises character-level word embeddings where each word is represented as a tensor of character indices (codes from the UNICODE table) with padding or cropping to a certain length. Two CNN layers with 1D kernels of different sizes are applied to the input tensors and the results are concatenated together.

Embeddings for all words from the medical report are summed together with a simple sum since it is independent of the word ordering in the document. It provides us with a single vector representation for each report.

The report representation is directly passed to a fully connected network that serves as a classification head. This model is hereafter called *DocChar*.

3.5 Hierarchical Attention GRU

As a representative of the recurrent neural network, we employ hierarchical attention GRU. This model should reflect the structure of a document since it contains a word and sentence encoder ((Yang et al., 2016)). Word embeddings are initialised randomly and are fed into a bidirectional GRU ((Cho et al., 2014)) layer, which is followed by the attention mechanism.

Our first intention was to use medical sections (paragraphs) instead of sentences. However, it is challenging to segment such unstructured reports into sections reliably. Therefore, we used fixed-length word sequences (50 sequences filled with 25 words). This model is hereafter called *HA-GRU*.

4 DATASET

The dataset consists of the medical reports collected between the years 2016 and 2021. The data were provided by a Czech hospital and are fully anonymised. A medical report contains several blocks, such as a diagnoses block containing descriptions of assigned diagnoses or a header with information about the hospital. However, the blocks are not structured, contain various headings and sometimes they are missing at all. Therefore, we consider the reports as unstructured text.

Example of input can be seen in Table 1.

All reports are annotated with a set of medical codes according to the ICD. The ICD-10 taxonomy contains more than 15,000 codes with a hierarchical structure. The first level grouping is according to chapters ranging from A to Z. We exclude chapters U to Y that are actually not diagnoses but some special purpose codes and external factors influencing the patient. The chapters are further divided into the ranges (e.g., F00-F99), sub-ranges (e.g., F00-F09), diagnoses (e.g., F01) and finally to the specific diagnoses (e.g., F01.1).

Due to the highly imbalanced numbers of the labels in the last level of the hierarchy, we have decided to concentrate on classifying the penultimate level codes (3 character codes, e.g., F01). According to the discussions with practitioners, the three character codes are sufficient for practical usage.

The dataset is divided into three parts: train 85%, test 10% and dev 5%. In the following text, we provide the statistics of the dataset. Figure 1 depicts the label co-occurrence matrix, which describes how frequently other related diagnoses occur if the examined diagnosis is present. It indicates that there are patterns of highly correlated diagnoses that can be utilised for model improvements as well as for validating the predictions.

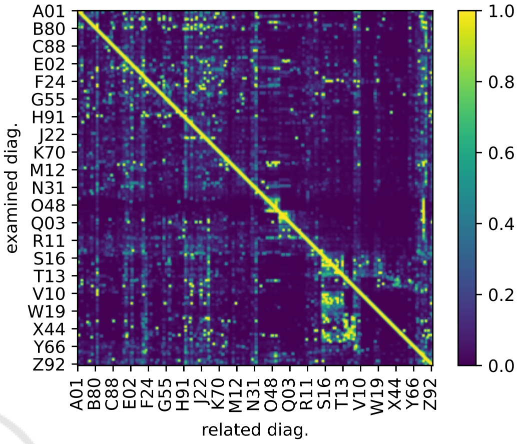

Figure 1: Label co-occurrence matrix (normalized and dilated by kernel 15x15 for visualization purposes).

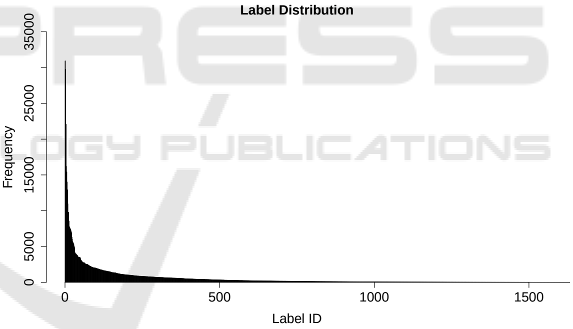

Figure 2: Label distribution in the dataset.

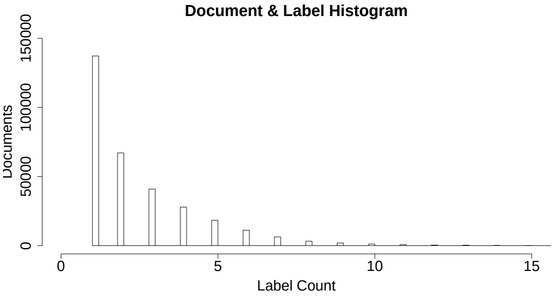

Figure 3: Label count histogram.

Figure 2 shows the distribution of the labels in the dataset. There are several very frequent classes and the distribution has a very long tail of rare diagnoses.

Figure 3 shows the label counts in particular docu-

Table 1: Example of a part of a message. left - czech, right - machine-translated to english.

ANAMNÉZA: OA: St.p. vysoké DVT vlevo v roce XX, stp. recid. PE, st.p. zavedení kaválního filtru, postrombotický syndrom vlevo, stav po úrazu kotníku vlevo, hematologicky dle vlastních slov vyšetřován nebyl a ani není v dispenzarizaci, v roce XXXX snad znovu tromboza LDK, trombosa VSM XXXX, lupus antikoagulans, homozygot PAI 4G/5G, zvýš. fVIII, pozit. proC global, chronicky warfarinizován, st.p. plastice kůže v oblasti kolene l.sin. po úrazu XXXX, jinak se s ničím neléčí, sledován jen u OL, kam chodí na kontroly INR. FA: Warfarin 3 mg 0-3-0 (nyní 5. den ex), Euphylin 300 1-0-1, Detralex 2-0-0, Vessel due 1-0-1. Abusus: 20 cig denně, alkohol: příležitostně - 10piv na posezení FF: v normě, váha 125 kg, výška 193 cm Alergie: neguje NYNĚJŠÍ ONEMOCNĚNÍ: Pacient s opakovanými trombosami hlubokého žilního systému DKK a recid. plicní embolizací se zavedeným kavafiltrem. Nyní na CT zjištěna nevhodná poloha filtru a nekompletní rozvinutí. Přijat k extrakci kavafiltru, ev. impl. Milesovy svorky.	Medical History: Previous high deep vein thrombosis (DVT) in the left leg Previous recurrent pulmonary embolism (PE) Cavain filter was inserted Post-thrombotic syndrome in the left leg Previous ankle injury in the left leg Hematological examination not performed according to patient's statement, not in dispense Possible DVT in the left leg in XXXX VSM thrombosis in XXXX Lupus anticoagulans Homozygous PAI 4G/5G Increased fVIII Positive proC global Chronic warfarin treatment Previous plastic surgery on the left knee after injury in XXXX Otherwise not receiving any treatment, only monitored by outpatient clinic for INR check-ups FA: Warfarin 3 mg 0-3-0 (currently on day 5 of treatment) Euphylin 300 1-0-1 Detralex 2-0-0 Vessel due 1-0-1 Abusus: 20 cigarettes per day Occasional alcohol consumption - 10 beers per social occasion FF: Within normal limits Weight: 125 kg Height: 193 cm CURRENT CONDITION: Patient with repeated DVT and recurrent PE, currently has a cavafilter inserted. CT showed that the filter is in an inappropriate position and incompletely deployed. Admitted for cavafilter extraction and possible implementation of Miles stenting.

ments. Almost 140,000 reports are labelled with only a single label determining the main diagnosis. Most of the documents then have up to 6 diagnoses in total and there are several reports with more diagnoses. Table 2 shows the statistics of the corpus. The values represent number of words within the dataset.

Table 2: Dataset statistics.

Part	All	Train	Test	Dev
Records	316,808	269,578	31,404	15,826
Avg text length	1,351	1,351	1,349	1,348
Avg label count	2.47	2.47	2.47	2.48

5 EXPERIMENTS

The performed experiments follow two scenarios. The first one, the main diagnosis classification (single-label), concentrates on determining the main diagnosis for each medical report. The main diagnosis is the most important one for the billing purposes and therefore, it is the priority for the target application.

The second scenario (multi-label classification) is to find all diagnoses, including the main one. The scenarios were defined in cooperation with practitioners who are supposed to use the outcomes of this study. In both scenarios, we deal with the 3 character codes as described in Section 4. The total number of labels is 1126 in the single-label scenario and 1523 in the multi-label one. The difference shows that not all diagnoses are used as the main one.

For the single-label classification, we report the accuracy and also the macro-averaged precision, recall and F-measure which takes into consideration the imbalanced label distribution. The multi-label classification is evaluated in terms of both micro- and macro-averaged precision, recall and F-measure.

5.1 Main Diagnosis Classification

Table 3 shows the comparison of the selected classification models for the *main diagnosis classification* scenario. Based on the results, we can state that most models perform comparably and slightly outperform the baseline approach. Best accuracies were obtained by CNN, Electra and HA-GRU. The differences among these models are very small and they are under the confidence level which is 0.5 % in our setting.

Table 3: Macro precision, recall, F-measure and accuracy for the *main diagnosis classification* scenario [in %].

Model	F1	P	R	Acc.
MLP (base.)	42.0	45.7	42.3	75.3
1-5 CNN 512	43.9	47.9	44.3	77.6
CNN 1024	43.8	48.8	43.6	78.0
ELECTRA	44.8	47.4	**46.2**	**78.3**
DocChar	44.5	**59.8**	45.5	74.8
HA-GRU	**45.1**	48.4	45.6	78.2

5.2 All Diagnoses Classification

This section deals with the multi-label *all diagnoses classification* scenario. The results are summarised in Table 4. We tested the same models as in the single-label scenario. The only modification is using the sigmoid activation function in the classification layer and binary cross-entropy loss function. In this scenario, HA-GRU is the best-performing model in terms of both micro- and macro-averaged values. The results indicate that the attention mechanism used in this network is the most suitable for the task and outperforms the more complex Electra model as well as the CNN networks.

Table 4: Macro- and micro-averaged precision, recall and F-measure for the *all diagnoses classification* scenario in [%].

Model	**Macro**			**Micro**		
	F1	**P**	**R**	**F1**	**P**	**R**
MLP (base.)	31.6	43.8	26.9	68.7	78.3	61.2
1-8 CNN 512	35.6	46.8	31.7	71.8	80.5	64.8
CNN 1024	34.2	43.9	31.3	72.0	81.3	64.4
ELECTRA	20.0	27.1	17.5	70.6	**83.3**	61.3
DocChar	33.9	47.1	29.6	65.2	80.5	54.8
HA-GRU	**41.8**	**50.3**	**38.3**	**75.1**	79.7	**71.1**

6 CONCLUSIONS AND FUTURE WORK

In this study, we have performed a comparative evaluation of several state-of-the-art models for the task of medical report classification in Czech. To the best of our knowledge, it is the first attempt at automatic diagnosis coding on Czech data.

The results for the main diagnosis scenario indicate that the models perform comparably and slightly outperform the baseline which proved to be relatively strong.

In the second scenario, the more sophisticated models obtained better results compared to the baseline. The HA-GRU model proved to be the best one in this scenario.

We can also conclude that the results of the best HA-GRU model are good enough to be integrated into the target system which will significantly reduce the workload of the operators and thus also saves the time and money.

In the future work, we would like to improve the architecture of the HA-GRU model and adjust it for utilisation of other types of clinical reports such as epicrisis etc. and improve the performance.

ACKNOWLEDGEMENTS

This work has been partly supported by Grant No. SGS-2022-016 Advanced methods of data processing and analysis.

REFERENCES

Baumel, T., Nassour-Kassis, J., Cohen, R., Elhadad, M., and Elhadad, N. (2018). Multi-label classification of patient notes: case study on icd code assignment. In *Workshops at the thirty-second AAAI conference on artificial intelligence*.

Chalkidis, I., Fergadiotis, M., Malakasiotis, P., Aletras, N., and Androutsopoulos, I. (2019). Extreme multi-label legal text classification: A case study in eu legislation. *arXiv preprint arXiv:1905.10892*.

Cho, K., Van Merriënboer, B., Bahdanau, D., and Bengio, Y. (2014). On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*.

Clark, K., Luong, M.-T., Le, Q. V., and Manning, C. D. (2020). ELECTRA: Pre-training text encoders as discriminators rather than generators. In *ICLR*.

El Boukkouri, H., Ferret, O., Lavergne, T., Noji, H., Zweigenbaum, P., and Tsujii, J. (2020). CharacterBERT: Reconciling ELMo and BERT for word-level open-vocabulary representations from characters. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6903–6915, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Gu, P., Yang, S., Li, Q., and Wang, J. (2021). Disease correlation enhanced attention network for icd coding. In *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1325–1330, Los Alamitos, CA, USA. IEEE Computer Society.

Johnson, A. E., Pollard, T. J., Shen, L., Lehman, L.-w. H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Anthony Celi, L., and Mark, R. G. (2016). Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9.

Kim, Y. (2014). Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*.

Kocián, M., Náplava, J., Štancl, D., and Kadlec, V. (2022). Siamese bert-based model for web search relevance ranking evaluated on a new czech dataset. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(11):12369–12377.

Lenc, L. and Král, P. (2016). Deep neural networks for czech multi-label document classification. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 460–471. Springer.

McAuley, J. and Leskovec, J. (2013). Hidden factors and hidden topics: understanding rating dimensions with review text. In *Proceedings of the 7th ACM conference on Recommender systems*, pages 165–172.

Mullenbach, J., Wiegreffe, S., Duke, J., Sun, J., and Eisenstein, J. (2018). Explainable prediction of medical codes from clinical text. *arXiv preprint arXiv:1802.05695*.

Perotte, A., Pivovarov, R., Natarajan, K., Weiskopf, N., Wood, F., and Elhadad, N. (2014). Diagnosis code assignment: models and evaluation metrics. *Journal of the American Medical Informatics Association*, 21(2):231–237.

Shi, H., Xie, P., Hu, Z., Zhang, M., and Xing, E. P. (2017). Towards automated icd coding using deep learning. *arXiv preprint arXiv:1711.04075*.

Sido, J., Pražák, O., Přibáň, P., Pašek, J., Seják, M., and Konopík, M. (2021). Czert – Czech BERT-like model for language representation. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1326–1338, Held Online. INCOMA Ltd.

Straka, M., Náplava, J., Straková, J., and Samuel, D. (2021). Robeczech: Czech roberta, a monolingual contextualized language representation model. *arXiv preprint arXiv:2105.11314*.

Vani, A., Jernite, Y., and Sontag, D. (2017). Grounded recurrent neural networks. *arXiv preprint arXiv:1705.08557*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. (2017). Attention is all you need. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Xun, G., Jha, K., Sun, J., and Zhang, A. (2020). Correlation networks for extreme multi-label text classification. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '20, page 1074–1082, New York, NY, USA. Association for Computing Machinery.

Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., and Hovy, E. (2016). Hierarchical attention networks for document classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 1480–1489.

You, R., Zhang, Z., Wang, Z., Dai, S., Mamitsuka, H., and Zhu, S. (2019). Attentionxml: Label tree-based attention-aware deep model for high-performance extreme multi-label text classification. *Advances in Neural Information Processing Systems*, 32.