Comparative Analysis of Fine-tuned Deep Learning Language Models for ICD-10 classification task for Bulgarian Language

Boris Velichkov[1,2], **Sylvia Vassileva**[1], **Simeon Gerginov**[1], **Boris Kraychev**[2], **Ivaylo Ivanov**[1],
Philip Ivanov[1], **Ivan Koychev**[1,2] and **Svetla Boytcheva**[3]
[1] Faculty of Mathematics and Informatics, Sofia University ”St. Kliment Ohridski”, Sofia, Bulgaria
sylvia.vassileva@gmail.com, simeongerginov1@gmail.com,
sashovi@uni-sofia.bg, fsivanov@uni-sofia.bg, koychev@fmi.uni-sofia.bg
[2] GATE Institute, Sofia University ”St. Kliment Ohridski”, Sofia, Bulgaria
boris.velichkov@gate-ai.eu, boris.kraychev@gate-ai.eu
[3] Institute of Information and Communication Technologies,
Bulgarian Academy of Sciences, Sofia, Bulgaria
svetla.boytcheva@iict.bas.bg

Abstract

The task of automatic diagnosis encoding into standard medical classifications and ontologies is of great importance in medicine - both to support the daily tasks of physicians in the preparation and reporting of clinical documentation, and for automatic processing of clinical reports. In this paper, we investigate the application and performance of different deep learning transformers for automatic encoding in ICD-10 of clinical texts in Bulgarian. The comparative analysis attempts to find which approach is more efficient to be used for fine-tuning of pre-trained BERT family transformer to deal with a specific domain terminology on a rare language such as Bulgarian. On the one hand, we use SlavicBERT and MultiligualBERT models, which are pre-trained for a common vocabulary in Bulgarian but lack medical terminology. On the other hand, we compare them to BioBERT, ClinicalBERT, SapBERT, BlueBERT models, which are pre-trained for medical terminology in English, but lack training for language models in Bulgarian, and vocabulary in Cyrillic. In our research study, all BERT models are fine-tuned with additional medical texts in Bulgarian and then applied to the classification task for encoding medical diagnoses in Bulgarian into ICD-10 codes. A big corpus of diagnoses in Bulgarian annotated with ICD-10 codes is used for the classification task. Such an analysis gives a good idea of which of the models would be suitable for tasks of a similar type and domain. The experiments and evaluation results show that both approaches have comparable accuracy.

1 Introduction

The task for automatic encoding of Electronic Health Records (EHR) with standard medical classifications is a hot-topic. The international classification of diseases, 10th revision (ICD-10)[1] is one of the most commonly used standard medical classifications due to the availability of translations in several languages. It is a hierarchical classification that encodes each diagnosis into a standard code which is used for statistical analysis and insurance reimbursement. The current solutions for this task are based on a restricted subset of ICD-10 codes or address some specific task trained on a small manually annotated corpus.

Recently some deep learning models like BERT (Devlin et al., 2018) pre-trained transformers were applied for different Natural Language Processing (NLP) and clinical NLP tasks. The first type of transformers is language models that cover common vocabulary on several languages: SlavicBERT (Arkhipov et al., 2019) and MultiligualBERT (Pires et al., 2019). The second type is transformers that cover specific terminology in English. In this particular case, the base language model is pre-trained with medical terminology by using scientific articles abstracts from PubMED[2] or full-text articles from PMC[3]: BioBERT (Lee et al., 2020), ClinicalBERT (Alsentzer et al., 2019), SapBERT (Liu et al., 2021), BlueBERT (Peng et al., 2019), MT-BERT (Peng et al., 2020) , PubMEdBERT (Gu et al., 2020). The later transformers prove that rel-

[1] https://icd.who.int/browse10/2019/en
[2] https://pubmed.ncbi.nlm.nih.gov/
[3] https://www.ncbi.nlm.nih.gov/pmc/

Proceedings of Recent Advances in Natural Language Processing, pages 1448–1454
Sep 1–3, 2021.
https://doi.org/10.26615/978-954-452-072-4_162

atively high accuracy can be achieved in training for automatic ICD-10 classification task for the English language (Moons et al., 2020). We hypothesise that comparable accuracy can be achieved also for languages other than English using either type of pre-trained BERT transformers.

2 Related Work

The task for automatic ICD-10 encoding of textual descriptions of diagnosis was addressed in several research challenges like i2b2 NLP Challenges, CLEF eHealth, etc. The major problem is that ICD-10 classification contains more than 11K codes and requires a significant number of labeled training data. In general, there are only available labeled datasets for a limited number of ICD-10 codes, which is one of the reasons why this task is not yet solved to the full range of ICD-10 codes. Lavergne et al presented in (Lavergne et al., 2016) a dataset for for ICD-10 coding of death certificates that contains 377,677 labeled statements with 3,457 unique ICD-10 codes. Usually, the labeled datasets are highly unbalanced that has a huge impact on the annotation method performance. This problem was addressed in (Parlak and Uysal, 2018), where the authors apply techniques for imbalance effects reduction, like splitting feature spaces and compressing label dimension. The ICD-10 classification task was investigated for several languages. The best performance for languages other than English was achieved with SVM models (Bagheri et al., 2020) F1 54.9% for Dutch; the longest common subsequence problem (Chen et al., 2017) for Chinese with F1-score of 81.1%; a hierarchical approach(Ning et al., 2016) for Chinese with F1 score of 91.08%; information retrieval techniques for Turkish (CEYLAN et al., 2012) with the best score of 76.5% Another approach is to view the problem as a multi-label classification task and use neural networks like CNN, LSTM/BiLSTM, and HA-GRU (Wang et al., 2020), or applying BERT which has shown good results on this task in German (Amin et al., 2019). Hybrid approaches, that combine different models show a slight improvement in the results (Amin et al., 2019).

For the Bulgarian language was done some preliminary experiments using SVM and small training corpora (Boytcheva, 2011), where the model achieved F-score 84%. We need to mention that the reported results in this work are based on significantly smaller training and test datasets with limited number of ICD-10 classes. In this paper we use big annotated corpora and include almost all ICD-10 codes used by medical practitioners in Bulgaria. In (Velichkov et al., 2020) we show some successful application of the BERT pretrained transformers for ICD-10 encoding. Inspired by the promising results we will investigate both language models: MultiligualBERT and SlavicBERT and will compare them with the state-of-the-art models for medical domain in English: BioBERT, ClinicalBERT, BlueBERT and SapBERT.

3 Data

3.1 Language Model Pre-training Dataset

For the Pre-training Dataset we have used a combination of medical articles and medical journals scraped from the internet. Medical articles were crawled from MedInfo[4]. We've also used a dataset of crawled medical articles that is already publicly available in GitHub[5]. Medical journals were crawled from MedUnion[6], JournalsMuVarna[7], MedicinaNauka[8], CmlMuSofia[9], Bulsem[10], Basa[11], MedSport[12] and Vma[13]. The crawled medical articles and journals were cleaned from single and double quotes, as well as any special characters and new lines.

From MedInfo we have crawled 1,740 medical articles. Each article describes different topics in terms of medical diseases, as well as possible treatments for the different diseases.

Each medical journal is in a PDF format and was split by page during crawling.

From MedUnion we have crawled 612 pages of medical journals. Each journal describes modern medicine and different aspects of it.

From JournalsMuVarna we have crawled 1,230 pages of medical journals. Each journal describes different topics like Social medicine, Health policy, Healthcare management, History of medicine and healthcare, and others.

[4] https://www.medinfo.bg/

[5] https://github.com/BorisVelichkov/scrapping-framar-and-bgmedic

[6] http://www.medunion-bg.org/

[7] https://journals.mu-varna.bg/index.php/sm/index

[8] https://medicina.nauka.bg/

[9] http://cml.mu-sofia.bg/CML/mpreg/index.html

[10] http://www.bulsem.bg/bg/about-jem

[11] https://www.basa.bg/

[12] https://www.med-sport.net/index.html

[13] https://www.vma.bg/

From MedicinaNauka we have crawled 281 pages of medical journals. Each journal consists of Bulgarian science and medicine topics and advises on how to tackle different medical issues that can occur.

From CmlMuSofia we have crawled 160 pages of medical journals. Each journal provides information about original scientific developments such as articles and reviews. Healthcare Management, Medical Ethics, and History of Medicine are also regularly covered in each journal.

From Bulsem we have crawled 1,924 pages of medical journals. Each journal has original articles from all fields of medicine and dentistry by Bulgarian and foreign authors.

From Basa we have crawled 1,353 pages of medical journals. Each journal consists of reviews, original articles, clinical cases, and case reports.

From MedSport we have crawled 1,793 pages of medical journals. Each journal covers problems of sports orthopedics, rehabilitation, physiology as well as the medical aspects of the training and competition process.

From Vma we have crawled 4,848 pages of medical journals. Each journal consists of scientific developments, publications from scientific medical forums, cases from the practice, and reports about new scientific events.

3.2 ICD-10 Classification Task Dataset

Table 1: ICD10 datasets statistics (only 4 sign codes).

Dataset	Total Inst.	Unique Codes	Inst. w. Altern. Codes	Unique Tokens
Full	354,733	5,879	55,372	79,732
Train	284,144	5,879	-	76,909
Dev	35,117	5,876	26,186	31,753
Test	35,472	5,861	29,186	31,958

Table 2: ICD10 datasets: descriptive statistics for the number of alternatives codes.

Dataset	Max	Mean	Median	Min
Full	24	1.421	1	1
Train	-	-	-	-
Dev	22	1.409	1	1
Test	24	1.431	1	1

Table 3: ICD10 datasets: descriptive statistics for the number of tokens.

Dataset	Max	Mean	Median	Min
Full	34	4.787	4.0	1
Train	34	4.785	4.0	1
Dev	30	4.793	4.0	1
Test	32	4.795	4.0	1

The ICD-10 classification contains several levels encoded with a different number of signs. The root level is encoded with the letters from the English alphabet, and subsequent levels append a number to the parent one. In this article we examine 3-sign and 4-sign codes, for example, the 3-sign *A00* is the code for *"Cholera"*, and 4-sign *A00.0* is encoding a specific type of cholera - *"Cholera due to Vibrio cholerae 01, biovar cholerae"*.

In the current article, we use the corpus[14] published by Boytcheva et al (Boytcheva et al., 2020) as a basis and we perform additional preprocessing. It consists of two datasets: one with 189,756 3-sign samples and the other with 383,042 4-sign samples. The unique codes (classes) for each dataset are 2,035 and 10,971, respectively. It is important to emphasize that the dataset with 4-sign codes also includes 3-sign codes. The descriptions are in Bulgarian, Latin, and transliterated from Latin to Cyrillic. The second dataset (containing 4-sign and 3-sign codes) is used to process and conduct experiments with different BERT family models. ICD-10 codes are numerous, with some having only a few samples in the dataset. In other words, the dataset is highly imbalanced. For this reason, additional processing has been done, which aims to achieve three things:

1. Add artificially created samples. This is done by applying the following data augmentation techniques: word exchange; exchange of random letters in one word; delete any letter of a word; change any letter in a word to one close to it on the keyboard.

2. Codes that have less than 5 samples should be reduced to a higher level in the code hierarchy. This is possible because ICD-10 codes have a strictly specific hierarchy. For example, a 4-sign code like A00.0 has 4 levels - each of its symbols. The highest level is the letter. The next three levels are the numbers. About 4,939 classes in the dataset have less than 5 samples.

[14] `https://github.com/BorisVelichkov/ICD10-Medical-Data`

Those with a 4-sign code are reduced to their corresponding 3-sign code (remove the 4-sign specific class from the classification). 405 classes with 3-signs are under-represented and thus we cannot apply this approach for them.

3. To unite the classes that are not related to a particular disease but have a special purpose for capturing external factors influencing health. These are the codes V01-Y98 (External causes of morbidity and mortality) and Z00-Z99 (Factors influencing health status and contact with health services). They are reduced to the upper levels V and Z, respectively, following ICD-10 grouping logic.

The converted dataset (Full) is divided into three parts: train (Train), validation (Dev), and test (Test) datasets. An additional column for alternative codes has been added to the validation and test datasets, as a diagnosis can be assigned to more than one code. For each dataset, the number of samples, the number of unique codes, the number of samples with alternative codes, and the number of unique tokens are shown in Table 1. It is good to note that in the validation set there are 1,708 unique tokens that are not present in the training set, as well as 1,701 tokens in the test set that are not present in the training set. The intersection between these tokens is approximately one-third - 586 common unique tokens. Descriptive statistics such as the minimum number, the maximum number, the mean, and the median of the alternative codes are shown in Table 2. The numbers vary between 1 and 24, with most being closer to only one alternative code. An equivalent table with descriptive characteristics for the number of tokens is Table 3. There, the number of tokens varies between 1 and 34, but the average is between 4 and 5.

4 Deep Learning Methods for text-based classification

BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2018) is a deep learning language model pre-trained on a large corpus of data using bidirectional transformers that provides context-aware token and sentence representations. There are multiple BERT models for different languages and domains and BERT has shown very good results on a variety of different tasks. Transfer learning can be applied by using the published models and fine-tuning them with a smaller dataset on the target task.

We evaluate multiple BERT models by applying additional pre-training for Bulgarian medical texts using the masked language task and then fine-tuning them on the multi-class classification for ICD-10 codes.

For the masked language task, we mask the standard 15% of tokens and train the model to predict the correct token following the architecture from the original paper (Devlin et al., 2018). The goal of training is to minimize the perplexity of the model. We use the language model pre-training dataset to improve BERT's understanding of Bulgarian medical text. We split the language model pre-training dataset in a proportion of 80:20 - 80% for training and 20% for testing.

WordPiece is used for tokenization and the original vocabulary from each model is used. As subword tokens are used, all words can be represented with tokens in the vocabulary. To train domain/language-specific extension to the vocabulary, a large corpus of training data would be required which is unavailable for Bulgarian.

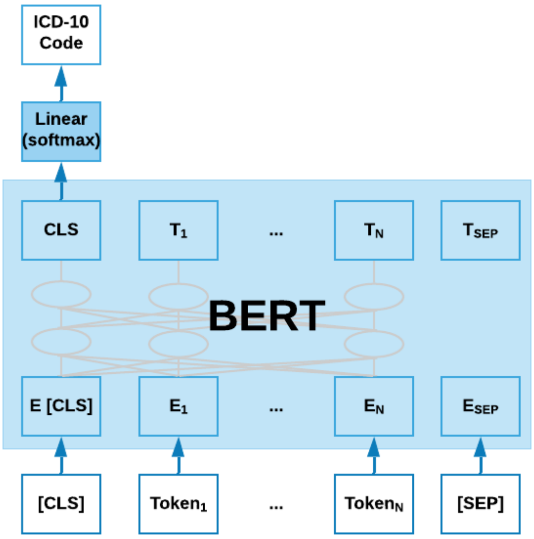

Figure 1: Architecture of the BERT text classifier.

When BERT tokenizes the input text, it prepends and appends two special tokens - [CLS] and [SEP]. The [CLS] token pools the information from all tokens in the sentence and in our case represents the diagnosis embedding which we use for text classification. For the multi-class classification task, we use the architecture proposed in (Devlin et al., 2018) (Fig. 2). We add a linear layer on top of BERT, which uses the [CLS] token output from the encoder and is trained to predict the correct ICD-10 class using a softmax activation. We return the top 5 classes with the highest probability as a prediction from the classifier as each diagnosis can belong to more than one class. We report accuracy, macro-F1, and mean reciprocal rank (MRR) metrics for the classification task.

The Multilingual BERT model uses BERT-base as a starting point and is additionally fine-tuned on the masked language task using Wikipedia articles in 104 languages incl. Bulgarian[15].

BioBERT is based on BERT-base and fine-tuned on PubMed abstracts and PMC full-text articles[16].

BlueBERT is a model based on BERT that is pre-trained on PubMed abstracts and (MIMIC-III[17]) clinical notes[18].

ClinicalBERT is initialized from BioBERT and trained on MIMIC-III, which contains around 2 million notes [19].

SapBERT is a PubMedBERT that was further fine-tuned with synonym pairs from the knowledge base of UMLS, a collection of biomedical ontologies[20].

SlavicBERT is a model, derived from MulitlingualBERT, trained on Wikipedia articles in Bulgarian, Czech and Polish and news in Russian [21].

5 Experiments and Results

Table 4: BERT fine-tuned models and their perplexity.

BERT model	Perplexity
BioBERT	1.7856
BlueBERT	1.8941
ClinicalBERT	1.7606
MultilingualBERT	3.2690
SapBERT	2.5644
SlavicBERT	5.6693

Table 5: Classification results for different BERT models.

BERT Model	Accuracy	Macro F1	MRR
BioBERT	78%	86%	91%
BlueBERT	71%	79%	87%
ClinicalBERT	92%	87%	94%
MultilingualBERT	87%	91%	95%
SapBERT	68%	76%	81%
SlavicBERT	90%	76%	93%

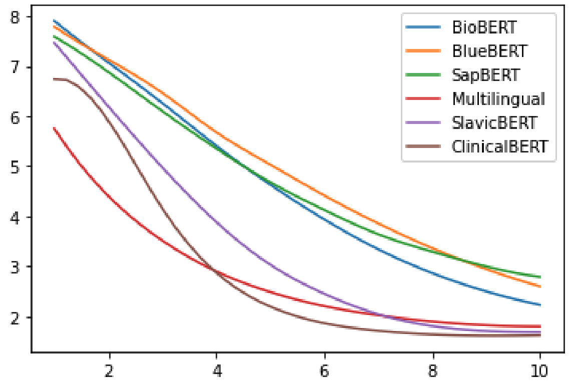

Figure 2: Evaluation loss per training epoch.

In the current article, experiments were performed with six different types of BERT models. Each was additionally fine-tuned on medical articles in Bulgarian and then attached to the classification task to associate the diagnosis with the corresponding ICD-10 code. ClinicalBERT is fine-tuned for 20 epochs with a final preplexity of 1.7606 (the 12th epoch was 1.7788). SlavicBERT is fine-tuned for 16 epochs and has a perplexity of 5.6693 (the 12th epoch was 5.7312). All other models are trained in 12 epochs. All perplexities can be seen in the Table 4. In the classification task, all models are trained in 10 epochs. The change in loss can be seen in Fig. 2. Detailed results including Accuracy, Macro F1 and MRR are shown in Table 5. It is noted that the highest MRR and Macro-F1 is using MultilingualBERT (95% and 91%, respectively), followed by ClinicalBERT with 1% below (94%) MRR and 87% Macro-F1. ClinicalBERT has the highest accuracy - 92%.

As we can see all the models are doing quite well. What makes the ClinicalBERT one of the best is that this model is pre-trained on top of many clinical notes, which contain a large amount of medical concepts. Many of them are in Latin and are the same in their use in different languages. Also these notes are most likely quite close to medical diagnoses. It is also the model that has been fine-tuned for most epochs (20 epochs versus 16 for the

[15] Multilingual BERT https://github.com/google-research/bert/blob/master/multilingual.md

[16] BioBERT https://github.com/dmis-lab/biobert

[17] MIMIC https://mimic.mit.edu/

[18] BlueBERThttps://github.com/ncbi-nlp/bluebert

[19] ClinicalBERT https://github.com/EmilyAlsentzer/clinicalBERT

[20] SapBERT https://github.com/cambridgeltl/sapbert

[21] SlavicBERT https://github.com/deepmipt/Slavic-BERT-NER

Table 6: MultilingualBERT and BioBERT models Top 5 predictions for 3 diagnosis of real patients.

Diagnosis Text	**Multilingual BERT**	**BioBERT**	True Class
"Захарен диабет 2 тип" (*Type 2 diabetes mellitus.*)	**E11**, E10, E12, P70.2, C91	E11, E10, E12, E13, N25.1	E11
"Хронична лимфоцитна левкемия, В-клетъчна, IV к.с. по Rai" (*Chronic lymphocytic leukemia, B-cell, IV hp according to Rai*)	C91, **C91.1**, C91.0, C83.5, C83	C91, C91.1, C91.0, C83.5, C94	C91.1
"Хронична лимфоцитна левкемия – В-кл., CD5+, IIIст. по Rai, «С» по Binnet CIRS score-16" (*Chronic lymphocytic leukemia - B-class, CD5 +, III st. by Rai, ”C” by Binnet CIRS score-16.*)	C94, C94.7, C88, C88.0, **C91**	C91, C91.0, C91.1, C83.5, C83	C91.1

SlavicBERT and 12 for the rest). This may be the reason why ClinicalBERT leads MultilingualBERT in the accuracy (with 5% better). MultilingualBERT, on the other hand, is trained in over 100 languages, which may allow it to do very well in different languages and to be relatively easy to be fine-tuned on new data. Similarly, it can be said that another advantage is that the diagnoses combine text in Bulgarian, Latin and transliterated from Latin to Cyrillic. In addition, we can say that in other studies for the same task in Bulgarian, MultilingualBERT is the model that gives the best results. Here it has best macro F1 and MRR.

In order to be able to illustrate in a more understandable way the task we will show three examples of real diagnoses from discharge letters of patients and how two of the classifiers (MultilingualBERT and BioBERT) predicted codes of these diagnoses in the Table 6. As we can see the two classifiers return the true code at first position for the first diagnose. Also for this sample BioBERT has more close predictions in the top 5. The second and the third diagnoses are a little bit more complex because they have 4-sign code which in these cases is same for both - *”C91.1”*. As can be seen, in addition to the same code, the two diagnoses are very similar in text. For both diagnoses, BioBERT returns the 3-sign code first and the exact 4-sign code second. The same thing is seen with MultilingualBERT, but only for the first of the two diagnoses. For the second, the results are worse and the classifier can only guess the 3-sign code of the diagnosis. An interesting observation is that it is also in fifth place in the top 5 predicted codes.

In contrast with the results presented in (Velichkov et al., 2020) the SlavicBERT model in our experiments shows comparable results with the other models, and moreover it is the second ranked model on the basis of accuracy. Another difference in our results is that ClinicalBERT outperforms BioBERT in all three evaluation metrics - accuracy, macro F1 and MRR. In both cases the reason is longer training (more epochs) for the fine-tuning of the models.

6 Conclusion

In the current article, a comparative analysis of six different BERT models is made, each of which is trained on a large amount of data and additionally fine-tuned on a big corpus of medical texts in Bulgarian. It can be said that the selected models are a good representative sample for the task on which they are applied, as among them there are models trained on over the top 100 languages, trained on Slavic languages, trained on medical and bio literature. The results obtained are quite high and show that all tested models are promising. As future improvements, it would be good for all models to be fine-tuned further, both with more texts and with more epochs. It would be good to equalize the number of epochs of fine-tuning for all models. Also other good improvements would be comparing the models before and after fine-tuning and applying cross-validation to more accurately evaluate the models.

Acknowledgments

This research is funded by the Bulgarian Ministry of Education and Science, grant DO1-200/2018 ’Electronic health care in Bulgaria’ (e-Zdrave).

Also funded via GATE project by the EU Horizon 2020 WIDESPREAD-2018-2020 TEAMING Phase 2 under GA No. 857155 and OP SE4SG under GA No. BG05M2OP001-1.003-0002-C01.

References

Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical bert embeddings. *arXiv preprint arXiv:1904.03323*.

Saadullah Amin, Günter Neumann, Katherine Dunfield, Anna Vechkaeva, Kathryn Chapman, and Morgan Wixted. 2019. Mlt-dfki at clef ehealth 2019: Multi-label classification of icd-10 codes with bert.

Mikhail Arkhipov, Maria Trofimova, Yurii Kuratov, and Alexey Sorokin. 2019. Tuning multilingual transformers for language-specific named entity recognition. In *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing*, pages 89–93.

Ayoub Bagheri, Arjan Sammani, Peter GM Van der Heijden, Folkert W Asselbergs, and Daniel L Oberski. 2020. Automatic icd-10 classification of diseases from dutch discharge letters. In *BIOINFORMATICS 2020-11th International Conference on Bioinformatics Models, Methods and Algorithms, Proceedings; Part of 13th International Joint Conference on Biomedical Engineering Systems and Technologies, BIOSTEC 2020*, volume 13, pages 281–289. SciTePress.

Svetla Boytcheva. 2011. Automatic matching of icd-10 codes to diagnoses in discharge letters. pages 11–18.

Svetla Boytcheva, Boris Velichkov, Gerasim Velchev, and Ivan Koychev. 2020. Automatic generation of annotated corpora of diagnoses with icd-10 codes based on open data and linked open data. In *2020 15th Conference on Computer Science and Information Systems (FedCSIS)*, pages 163–167. IEEE.

Nefise Meltem CEYLAN, Adil ALPKOÇAK, and Afsun Ezel ESATOĞLU. 2012. Tıbbi kayıtlara icd-10 hastalık kodlarının atanmasına yardımcı akıllı bir sistem.

YunZhi Chen, HuiJuan Lu, and LanJuan Li. 2017. Automatic icd-10 coding algorithm using an improved longest common subsequence based on semantic similarity. *PloS one*, 12(3):e0173410.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2020. Domain-specific language model pretraining for biomedical natural language processing. *arXiv preprint arXiv:2007.15779*.

Thomas Lavergne, Aurélie Névéol, Aude Robert, Cyril Grouin, Grégoire Rey, and Pierre Zweigenbaum. 2016. A dataset for icd-10 coding of death certificates: Creation and usage. In *Proceedings of the Fifth Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM2016)*, pages 60–69.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Fangyu Liu, Ehsan Shareghi, Zaiqiao Meng, Marco Basaldella, and Nigel Collier. 2021. Self-alignment pretraining for biomedical entity representations. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4228–4238.

Elias Moons, Aditya Khanna, Abbas Akkasi, and Marie-Francine Moens. 2020. A comparison of deep learning methods for icd coding of clinical records. *Applied Sciences*, 10(15):5262.

Wenxin Ning, Ming Yu, and Runtong Zhang. 2016. A hierarchical method to automatically encode chinese diagnoses through semantic similarity estimation. *BMC medical informatics and decision making*, 16(1):1–12.

Bekir Parlak and Alper Kursat Uysal. 2018. On feature weighting and selection for medical document classification. In *Developments and Advances in Intelligent Systems and Applications*, pages 269–282. Springer.

Yifan Peng, Qingyu Chen, and Zhiyong Lu. 2020. An empirical study of multi-task learning on bert for biomedical text mining. *arXiv preprint arXiv:2005.02799*.

Yifan Peng, Shankai Yan, and Zhiyong Lu. 2019. Transfer learning in biomedical natural language processing: An evaluation of bert and elmo on ten benchmarking datasets. In *Proceedings of the 2019 Workshop on Biomedical Natural Language Processing (BioNLP 2019)*, pages 58–65.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual bert? *arXiv preprint arXiv:1906.01502*.

Boris Velichkov, Simeon Gerginov, Panayot Panayotov, Sylvia Vassileva, Gerasim Velchev, Ivan Koychev, and Svetla Boytcheva. 2020. Automatic icd-10 codes association to diagnosis: Bulgarian case. In *CSBio'20: Proceedings of the Eleventh International Conference on Computational Systems-Biology and Bioinformatics*, pages 46–53.

Qiong Wang, Zongcheng Ji, Jingqi Wang, Stephen Wu, Weiyan Lin, Wenzhen Li, Li Ke, Guohong Xiao, Qing Jiang, Hua Xu, and Yi Zhou. 2020. A study of entity-linking methods for normalizing chinese diagnosis and procedure terms to icd codes. *Journal of Biomedical Informatics*, 105:103418.