Application of Deep Learning Methods to SNOMED CT Encoding of Clinical Texts: From Data Collection to Extreme Multi-Label Text-Based Classification

Anton Hristov[1,2], **Aleksandar Tahchiev**[1], **Hristo Papazov**[3], **Nikola Tulechki**[1], **Todor Primov**[1] and **Svetla Boytcheva** [1,4]
[1]Sirma AI (Ontotext), Bulgaria
anton.hristov@ontotext.com, alexander.tahchiev@ontotext.com, nikola.tulechki@ontotext.com, todor.primov@ontotext.com, svetla.boytcheva@ontotext.com
[2] Faculty of Mathematics and Informatics, Sofia University "St. Kliment Ohridski", Bulgaria
[3] MIT, Mathematics PhD Student, USA
h.g.papazov@gmail.com
[4] Institute of Information and Communication Technologies, Bulgarian Academy of Sciences, Bulgaria

Abstract

Concept normalization of clinical texts to standard medical classifications and ontologies is a task with high importance for healthcare and medical research. We attempt to solve this problem through automatic SNOMED CT encoding, where SNOMED CT is one of the most widely used and comprehensive clinical term ontologies. Applying basic Deep Learning models, however, leads to undesirable results due to the unbalanced nature of the data and the extreme number of classes. We propose a classification procedure that features a multiple-step workflow consisting of label clustering, multi-cluster classification, and clusters-to-labels mapping. For multi-cluster classification, BioBERT is fine-tuned over our custom dataset. The clusters-to-labels mapping is carried out by a one-vs-all classifier (SVC) applied to every single cluster. We also present the steps for automatic dataset generation of textual descriptions annotated with SNOMED CT codes based on public data and linked open data. In order to cope with the problem that our dataset is highly unbalanced, some data augmentation methods are applied. The results from the conducted experiments show high accuracy and reliability of our approach for prediction of SNOMED CT codes relevant to a clinical text.

1 Introduction

The task of automatic encoding of clinical text with standard medical classifications and ontologies is with high importance for healthcare organizations and medical research. Truly, more than 80% of clinical documents are stored in free-text format. This paper presents a research effort in solving the problem of automatic encoding of textual description of medical diagnoses with one of the most widely used (Lee et al., 2013) and comprehensive ontologies – the Systematized Nomenclature of Medicine – Clinical Terms (SNOMED CT)[1]. One of the most important characteristics of SNOMED CT, which makes it significantly different from the rest of the standard medical classifications, is that it is based on compositional grammar[2]. Another aspect of popularity and importance of SNOMED CT for health information is interoperability, that is discussed in (Peterson and Liu, 2020). SNOMED CT is well known for being one of the most comprehensive medical ontologies, which makes the task of automatic encoding an extreme scale classification task with more than 360,000 medical codes. Currently, this task has not been solved with sufficient accuracy for all possible classes. Usually the developed solutions cover restricted terminology from 10 to a couple of thousands terms (Gaudet-Blavignac et al., 2021). The compositional nature of the SNOMED CT codes gives us the opportunity to address the problem either with classical approaches for classification tasks or with specific solutions that benefit from the compositional grammar's structure. As a product of our research,

[1]https://www.snomed.org/
[2]https://confluence.ihtsdotools.org/display/DOCSCG/Compositional+Grammar+-+Specification+and+Guide

Proceedings of Recent Advances in Natural Language Processing, pages 557–565
Sep 1–3, 2021.
https://doi.org/10.26615/978-954-452-072-4_063

the developed service for automatic encoding with SNOMED CT codes will be used mainly for Electronic Health Records (EHR) processing for patients with oncological diseases and certain rare diseases. Most of these diseases are well known to have a huge number of related diseases. Thus, our study will not be restricted only to the diseases of interest but will have a much broader scope. We propose an adaptation of the approach proposed by (Chang et al., 2020) and demonstrate the entire process from training dataset construction to classification model design and training.

2 Related Work

The problem of automatic encoding of EHR with SNOMED CT codes was investigated by many researchers since the very beginning of the ontology development. Different solutions cover broad range of SNOMED CT codes from 10 to a couple of thousands, usually the main obstacle for scalabily is the availability of sufficient volume of annotated training data. Basaldella et al (Basaldella et al., 2020) present COMETA - manually annotated corpora by experts that contain 20k English biomedical entities encoded with SNOMED CT.

The most popular approaches for automatic encoding, include hybrid methods combining regular expressions and vector space models (Ruch et al., 2008) with top precision 0.823 and mean avg precision 0.45 for 1239 MEDLINE citations. Some approaches take in consideration compositional structure (Liu et al., 2012) of SNOMED CT.

Recent research is based on deep learning techniques, and the most promising solutions are using transformers like BERT (Devlin et al., 2018). Pattisapu et al (Pattisapu et al., 2020) apply word embeddings, graph embeddings and BERT derivatives transformers and achieve the highest accuracy 0.83 for two benchmark datasets CADEC and PsyTAR.

Kraljevic et al (Kraljevic et al., 2021) propose MedCAT with Macro F1: 0.841–0.860 across different clinical domains and tasks. MedCAT is based on Word2Vec embeddings, and there is also MedCAT BERT version based on clinicalBERT (Alsentzer et al., 2019), and latter model shows a little bit lower performance than the former one.

A recent systematic review (Gaudet-Blavignac et al., 2021) shows that only few of the developed services for automatic encoding with SNOMED CT, are provided as open source - The clinical Text Analysis and Knowledge Extraction System (cTAKES) (Savova et al., 2010) and MetaMap (Aronson, 2001) . Both of them are rule-based.

3 Data

One of the key factors that plays a role in the automatic encoding of SNOMED CT codes is the data. In our project, we do not have annotated data which can be used to train the developed models. Thus, we use certain public data and linked open data in order to automatically generate annotated corpora that can serve as a training dataset.

3.1 Data Sources

In our research, we will consider only a subset of the available SNOMED CT codes, namely those related to disorders, clinical findings and procedures. The relevant medical ontologies, standard classifications, and vocabularies for the project, which are used to enrich the SNOMED CT descriptions with additional alternative textual descriptions, are the Human Disease Ontology[3], the International Classification of Diseases, 10th revision (ICD-10)[4], the International Classification of Diseases, 9th revision (ICD-9)[5], the International Classification of Diseases for Oncology, 3rd Edition (ICD-O-3)[6], the Medical Subject Heading (MESH)[7], the Mondo Disease Ontology (MONDO)[8], the Orphanet Rare Disease Ontology (ORDO)[9], and the Unified Medical Language System (UMLS)[10]. Benefiting from the resources provided by the linked open data cloud (LOD)[11], we can identify some of the mappings between the ontologies listed above using Bioportal[12]. Some general equivalence mappings are provided for ICD-10 and ICD-9[13] as well as rules for mappings between SNOMED CT and ICD-10[14]. In addition, the ICD-10 CM Alphabeti-

[3] https://disease-ontology.org/
[4] https://icd.who.int/browse10/2019/en
[5] https://apps.who.int/iris/handle/10665/39473
[6] http://apps.who.int/iris/bitstream/handle/10665/96612/9789241548496_eng.pdf
[7] https://www.ncbi.nlm.nih.gov/mesh/
[8] https://mondo.monarchinitiative.org/
[9] https://www.orpha.net/consor/cgi-bin/index.php
[10] https://www.nlm.nih.gov/research/umls/index.html
[11] https://lod-cloud.net/
[12] https://bioportal.bioontology.org/
[13] https://shorturl.at/vIKP4
[14] https://www.nlm.nih.gov/research/umls/mapping_projects/snomedct_to_icd10cm.html

cal Index[15] is used. All these resource and official mappings between them provide valid encoding of the textual descriptions of diseases with SNOMED CT codes.

3.2 Data Integration

As we alluded above, in order to increase the size of the corpus of SNOMED codes with textual descriptions (and hence boost the predictive power of our neural network by ensuring a richer training set), we extract mappings between SNOMED CT and other medical ontologies, standard classifications, and vocabularies. Then, we use these mappings to link SNOMED CT codes to descriptions native to the aforementioned resources.

Following this guiding principle, but applying it to different subsets of SNOMED CT codes and allowing different degrees of *description transitivity* (stay tuned), we constructed four distinct datasets, the last two of which allowed for a high classification accuracy.

	Corpus Size	SNOMED CT Codes	Unique Descriptions
Dataset V1	22M	128k	280k
Dataset V2:	626k	227k	469k
Procedures	106k	64k	105k
Findings	140k	65k	107k
Disorders	380k	98k	257k
Dataset V3	85k	14k	54k
Dataset V4	198k	14k	58k

Table 1: Dataset Evolution

We can conceptualize the overarching principle behind the construction of the different datasets as follows: First, we choose a certain subset of SNOMED CT codes whose elements will serve as labels in the classification procedure. Second, we consider the medical codes from the above ontologies, classifications, and vocabularies that are linked to our chosen SNOMED CT subset through an "exact-match" type predicate.[16] Third, we build a graph whose vertices are all of the chosen SNOMED CT codes and their "exact-match" neighbors; and whose edges are precisely these "exact-match" mappings. Fourth, we prescribe a degree of description transitivity. That is, we specify whether medical codes in connected components will share all textual descriptions associated with that component or simply the descriptions associated with their immediate neighbors. Finally, we extract a corpus of SNOMED CT codes along with the natural language descriptions that these codes acquired from the mapping graph.

Version 1 of our dataset reflected the naïve idea of considering a graph with full description transitivity. Of course, this approach of total transitive search between ontologies is largely misguided since the mappings between SNOMED CT codes an outside resources are rarely one-to-one. Indeed, these mappings prescribe similarity rather than identity – a circumstance that caused the graph generated by the V1 SNOMED CT subset to contain a connected component encompassing more than 90% of the vertices. Thus, the majority of the relevant SNOMED CT codes became indistinguishable for our classification procedure.

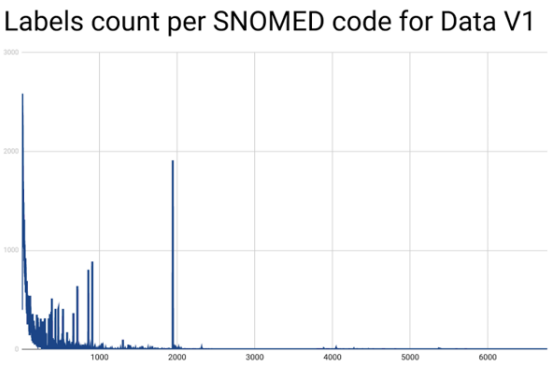


Figure 1: Dataset V1

For the second version of our dataset, we considered an even larger subset of SNOMED CT codes, which we further split into three categories – Procedures, Findings, and Disorders – respecting the official SNOMED CT organization. We used these three subsets to generate three distinct mapping graphs following the above procedure. Again, we prescribed full mapping transitivity on the individual graphs, but we forbade communication between graphs. This new strategy decreased the edge density of the graphs considerably, but large connected components were still present, which led to poor classification performance.

Consequently, we decided to narrow down the generating subset of SNOMED CT codes and to

[15] https://icd.codes/icd10cm/alphabetical-index

[16] We used verified existing mappings between ontologies and SPARQL queries (e.g. https://w.wiki/3ZXd) to extract similarities.

allow description-sharing only among immediate "exact-match" neighbors. Thus, for Dataset V3, we considered solely the SNOMED CT codes which exactly matched the following widely encountered procedures, findings, and disorders related to oncological diseases, certain rare diseases, and digestive, neurological, and respiratory diseases.

The lack of description transitivity in V3 caused the majority of the considered SNOMED CT codes to have unique description clusters. This circumstance allowed us to observe high classification accuracy for the first time. However, many SNOMED CT codes were now matched to a single single-word textual description, and the augmentation strategies discussed in the next subsection failed to meaningfully increase the description clusters of such codes.

For that reason, the official verified SNOMED CT ontology mappings used in V3 were supplemented with additional mappings excavated from Wikidata[17]. Then, full description transitivity was applied, and thus Dataset V4 came to be.

	Corpus Size	SNOMED CT Codes	Unique Descriptions
Additional Data	112k	8k	3k

Table 2: Additional Data

3.3 Data Augmentation

Our data integration strategies resulted in a one-to-many mapping of SNOMED CT codes to synonymous natural language descriptions. This mapping, however, featured a significant number of SNOMED CT codes with less than four textual descriptions. In order to address this circumstance, which would have otherwise interfered with the precision of our neural network, we employed several data augmentation techniques aimed at synthetically increasing the set of descriptions so that each SNOMED CT code could get mapped to at least four descriptions.

Our augmentation strategies were motivated by considerations of what synonymous textual descriptions could arise in the work of medical professionals.

[17] https://www.wikidata.org/wiki/Wikidata:Main_Page

3.3.1 Random Swap and Random Synonym Insertion

We adapted some of the code developed for *Easy Data Augmentation*(Wei and Zou, 2019) for the purposes of random word swapping and random synonym insertion. The *Random Swap* transformation works by selecting two random indices in a list of multiple words, and then, swapping the words with the corresponding indices. Only one swap is performed per transformation, which guarantees that novel descriptions are produced after the augmentation. The *Random Synonym Insertion* transformation works by shuffling the words in a sentence and looping over the shuffled sequence of words until a word with a WordNet[18] synonym is selected. Once such a word is found, a random synonym is pulled from its list of synonyms and inserted at a random place in the initial sentence. If no synonyms are found, nothing happens.

Examples:

- **Random Swap:** Fear of thunderstorms → Fear thunderstorms of.
- **Random Synonym Insertion:** Complete loss of teeth due to trauma → Complete loss of hurt teeth due to trauma.

3.3.2 Typographical Augmentation

Since medics work in tense environments, they are susceptible to making errors while typing medical records, as they are subjected to a lot of stress, strain, and lack of sleep. We have developed the following augmentations mimicking potential typos:

- **Swap adjacent character:**
  Syndrome → Synrdome.
- **Remove character:**
  Syndrome → Syndome.
- **Change character with corresponding adjacent keyboard-key character:**
  Syndrome → Syndrone.
- **Insert adjacent keyboard-key character to a word:**
  Syndrome → Syundrome.

All these augmentations are applied on a randomly selected character of a randomly selected word.

[18] https://wordnet.princeton.edu/

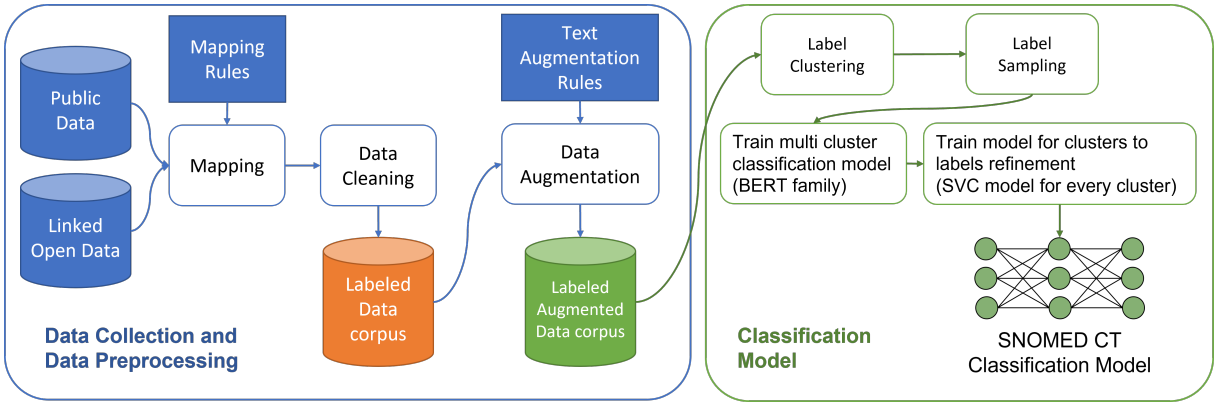


Figure 2: Data collection and classification model pipeline

3.3.3 Manual Augmentation

At certain places, where the above strategies could not be naturally applied, synonymous natural language descriptions were manually crafted.

Example:
3-PGDH deficiency → 3-phosphoglycerate dehydrogenase deficiency.

4 Text-Based Encoding with SNOMED CT Codes

If we have an unbalanced dataset, or if we want to split our problem into sub-problems, we can group our labels into clusters and train a model to predict to which of the clusters each sample belongs. After that, another model can refine (map) every predicted cluster to a specific label.

The proposed approach of text-based SNOMED CT classification is the following (see Fig. 2):

- Data Augmentation
- Label Clustering
- Sampling
- Train Multi-Cluster Classification Model
- Train Model for Clusters to Labels Refinement

4.1 Data Augmentation

The data augmentation techniques used in this step are described in detail in the previous section. The following parameters are used:

$$\begin{aligned} \text{augment probability} &= 40\% \\ \text{minimum samples} &= 5, \end{aligned}$$

where augment probability refers to what portion of the current description's words will be augmented and minimum samples refers to the minimum number of description samples every class should contain after the augmentation.

4.2 Label Clustering

Label clustering is widely used in extreme scale classification problems because datasets are mainly unbalanced and there are vast numbers of classes or because we want the classification task to be performed with less granularity. Our approach for clustering the dataset labels into groups is done by label embeddings, used in (Khandagale et al., 2020), and by applying clustering algorithm to it. Label embeddings for specific labels can be produced by summing all sample embeddings for which it is active. So, if we denote X to be the matrix holding the embeddings of all samples' descriptions, $X = [\text{samples} \times \text{embeddings}]$, and Y to be the matrix holding multi-hot encoding of samples' classes, $Y = [\text{samples} \times \text{labels}]$, Label embeddings (matrix L) is calculated using dot product between X and Y. This matrix L gives us information on how each label relates to each sample in our data, $L = Y^T X = [\text{labels} \times \text{embeddings}]$. For encoding the input samples' descriptions, we applied the pretrained BioBERT model (Lee et al., 2020). Clustering is done by a K-Means algorithm, with selected number of clusters of 100. This specific number is selected by manual analysis of data distribution over different number of clusters. The desired number of clusters is the smallest number that produces the minimum number of labels contained in more than one cluster, as well as best distribution of the labels for each cluster.

4.3 Sampling

Since our dataset is highly unbalanced because of the specifics of the domain, there are classes with only 5 samples and there are classes with more than 800 samples. The naïve sampling of random data splitting to train/dev/test sets will not work because this will result in all samples of some classes to be included entirely into one of the splits, which will lead to reduced accuracy. So, we developed a custom sampling strategy, which extracts distributed number of samples by a random manner into the dev and test split, which will result specific classes included into the dev and test splits to be included into the training corpus. For example, if class N has only 5 samples, its samples will be distributed as follows: train/dev/test = 3/1/1.

4.4 Train Multi-Cluster Classification Model

After label clustering is applied on the data, the next step is training a model for classification of the produced clusters. This can be formulated as binary, multi-class or multi-label classification. Since our dataset is very complex, and one class can be included into one or multiple clusters, we modified the official BioBERT implementation to perform a multi-label classification using Area Under the Curve (ROC AUC) as a scoring function. As an input, we have transformed our dataset using the produced clusters instead of the original labels, and we have fine-tuned the BioBERT weights on it.

4.5 Train Model for Clusters to Labels Refinement

The goal of the last step in the pipeline is finding to which label every sample belongs based on the already predicted cluster. There are a lot of possible solutions for mapping the sample's predicted clusters to their labels, and a classical one is for each label to look into all instances which is too expensive, discussed in (Chang et al., 2020). Another approach is training multiple models. A model for every label including the subset of all instances included into the predicted clusters (Chang et al., 2020), which will lead to the number of models equal to the number of classes. Since we are dealing with an extreme scale classification task with more than 10k classes, we think that this is not practical for applying it in real applications. We have trained liner one-vs-rest classifiers (Support Vector Classification (Platt, 1999), (Chang and Lin, 2011)) for every cluster including all its instances with BioBERT sample embeddings as an input. This approach results in 100 SVC trained models for clusters-to-labels refinement using Area Under the Curve (ROC AUC) as a scoring function.

Our contribution:

- Proposed augmentation techniques matching data distribution specifics;
- Proposed sampling strategy dealing with unbalanced data;
- Multi-class cluster classification is replaced with multi-label cluster classification, increasing the task's level of complexity;
- Cluster to label refinement is compressed to model per cluster, which is more suitable for extreme scale classification tasks;

5 Experiments and Results

5.1 Dataset V1

On version 1 of our dataset, we initially attempted a classical multi-class classification approach by using pretrained BioBERT (Lee et al., 2020). The results were close to random guessing, so we tried another approach. We used some standard community detection algorithms (like Louvain (Blondel et al., 2008), (Dugué and Perez, 2015), (Traag et al., 2011) and Leiden (Traag et al., 2019)) to group SNOMED CT codes into classes in order to train a cascade of hierarchical BioBERT classifiers. This grouping was necessary because 95% of Dataset V1 forms a dense graph (see Figure 5). After analyzing the results, we concluded that by this grouping a lot of important connections were removed. For this reason, we have left this approach aside.

5.2 Dataset V2

On the new version of the dataset, we tried to solve the problem by a multi-label classification approach using pretrained BioBERT again. After comprehensive training iterations, our model reached Area Under the Curve (ROC AUC) of 0.60 (Figure 6) which was not high enough for solving the problem.

5.3 Dataset V3

The proposed approach described in section 4 is applied on this third version of our dataset. We have fine-tuned the BioBERT weights for the multi-cluster classification task (Step 4 of our pipeline),

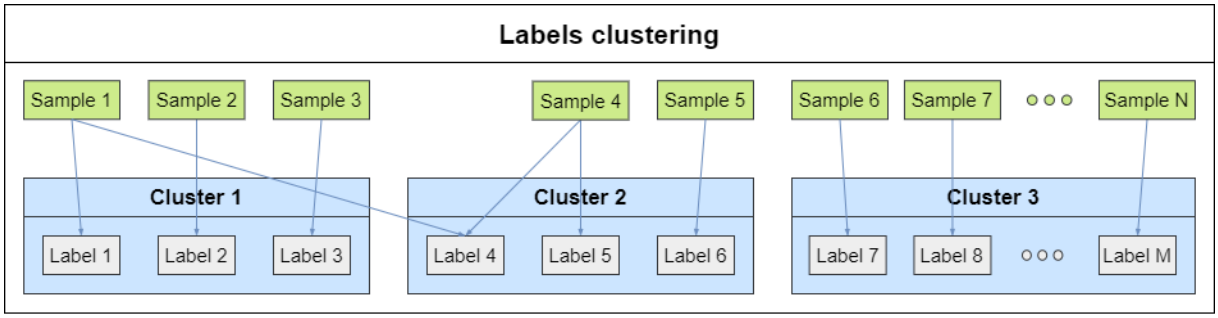


Figure 3: Cluster classification BioBERT model is dealing with multi-label data

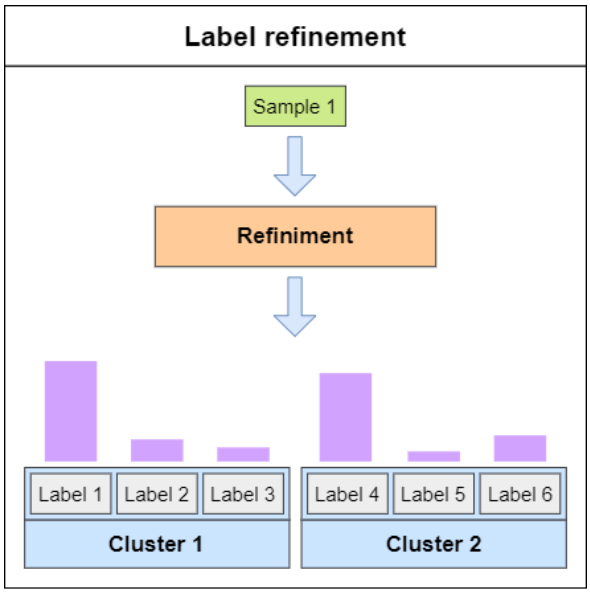
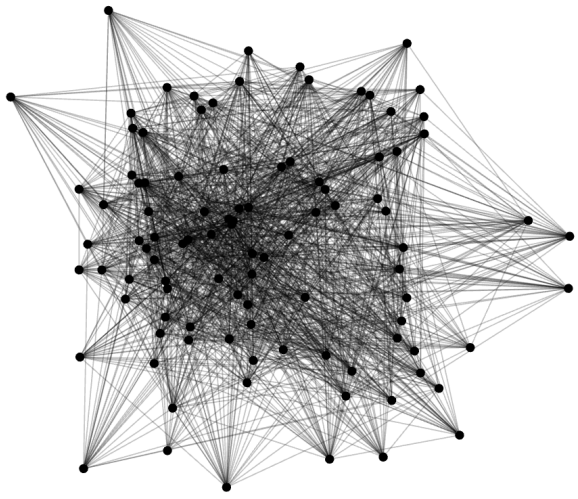


Figure 4: Refinement maps every sample with its corresponding labels based on the sample's predicted cluster



95% of Data V1 connections

Figure 5: Dataset V1 - Dense Graph

and after 7 epochs of training, it reached Area Under the Curve (ROC AUC) of 0.99653. Training

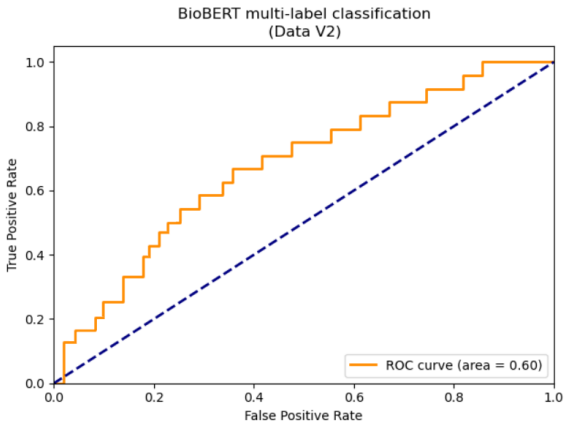


Figure 6: Dataset V2 - ROC curve

SVC for every cluster (100 SVC models) produces Area Under the Curve (ROC AUC) of 0.83273.

5.4 Dataset V4

The proposed approach described in the previous section is applied on this fourth version of our dataset. The Vocabulary is based on BioBERT v1.1, trained over PubMed. Some of the characteristics of Dataset V4 are presented in Table 3.

Characteristic	Count
Unique tokens	6,522
Number of tokens	2,949,353
Min of tokens	1
Max of tokens	189
Mean of tokens	14.85
Median of tokens	11.0

Table 3: Dataset V4 Characteristics

We have fine-tuned the BioBERT weights for the multi-cluster classification task (Step 4 of our pipeline), and after 25 epochs of training, it reached Area Under the Curve (ROC AUC) of 0.977.

Training SVC for every cluster (100 SVC mod-

els) produces Area Under the Curve (ROC AUC) of 0.804.

5.5 Discussion

The comparison (see Table 4) of the bioBERT multi-label classification with bioBERT clustering and label refinement, show that the proposed approach significantly improves the accuracy for SNOMED CT encoding task. The experiments are performed for trained models for 7 epochs for Dataset V2 - disorders subset.

Approach	Accuracy
bioBERT multilabel classification	0.56
bioBERT clusterings + label refinement (final approach)	0.97

Table 4: Comparison of the bioBERT models for Dataset V2- disorders

The proposed approach shows high accuracy and scalability. The additional steps for label refinement do not cause significant slow down of the learning process of the model. The obtained accuracy of the given method shows a significant improvement of the evaluation, compared to other solutions in the literature of the same problem, that report accuracy in the range from 0.83 (Pattisapu et al., 2020) up to 0.86 (Kraljevic et al., 2021) . Moreover, the presented results of the experiments and the evaluation are for a larger dataset and a wider range of SNOMED CT codes.

6 Conclusion and Further Work

We demonstrated how can be generated annotated dataset with SNOMED CT codes. The proposed approach demonstrates high accuracy and scalability. In comparison with other state-of-the-art approaches the achieved accuracy for the proposed model is relatively high and more over for wider coverage of SNOMED CT.

Our further work includes training of Multilingual BERT to solve the multilingual problem. Possible increase of the Area Under the Curve (ROC AUC) scores can be achieved through grid search applied to the selection of the K-Means clusters number, until finding the optimal number, based on the distribution of the data.

Acknowledgments

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825292 (ExaMode, http://www.examode.eu/).

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*.

Alan R Aronson. 2001. Effective mapping of biomedical text to the umls metathesaurus: the metamap program. In *Proceedings of the AMIA Symposium*, page 17. American Medical Informatics Association.

Marco Basaldella, Fangyu Liu, Ehsan Shareghi, and Nigel Collier. 2020. Cometa: A corpus for medical entity linking in the social media. *arXiv preprint arXiv:2010.03295*.

Vincent D. Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008.

Chih-Chung Chang and Chih-Jen Lin. 2011. Libsvm: A library for support vector machines. 2(3).

Wei-Cheng Chang, Hsiang-Fu Yu, Kai Zhong, Yiming Yang, and Inderjit S. Dhillon. 2020. Taming pretrained transformers for extreme multi-label text classification. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD'20, pages 3163—-3171. Association for Computing Machinery.

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*.

Nicolas Dugué and Anthony Perez. 2015. Directed Louvain : maximizing modularity in directed networks. Research report, Université d'Orléans.

Christophe Gaudet-Blavignac, Vasiliki Foufi, Mina Bjelogrlic, and Christian Lovis. 2021. Use of the systematized nomenclature of medicine clinical terms (snomed ct) for processing free text in health care: systematic scoping review. *Journal of medical Internet research*, 23(1).

Sujay Khandagale, Han Xiao, and Rohit Babbar. 2020. Bonsai: diverse and shallow trees for extreme multi-label classification. *Machine Learning*, 109(11):2099–2119.

Zeljko Kraljevic, Thomas Searle, Anthony Shek, Lukasz Roguski, Kawsar Noor, Daniel Bean, Aurelie Mascio, Leilei Zhu, Amos A Folarin, Angus Roberts, et al. 2021. Multi-domain clinical natural language processing with medcat: the medical concept annotation toolkit. *Artificial Intelligence in Medicine*, 117.

Dennis Lee, Ronald Cornet, Francis Lau, and Nicolette De Keizer. 2013. A survey of snomed ct implementations. *Journal of biomedical informatics*, 46(1):87–96.

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.

Hongfang Liu, Kavishwar Wagholikar, and Stephen Tze-Inn Wu. 2012. Using snomed-ct to encode summary level data–a corpus analysis. *AMIA Summits on Translational Science Proceedings*, 2012.

Nikhil Pattisapu, Sangameshwar Patil, Girish Palshikar, and Vasudeva Varma. 2020. Medical concept normalization by encoding target knowledge. In *Machine Learning for Health Workshop*, pages 246–259. PMLR.

Kevin J Peterson and Hongfang Liu. 2020. Automating the transformation of free-text clinical problems into snomed ct expressions. *AMIA Summits on Translational Science Proceedings*, 2020.

John C. Platt. 1999. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *ADVANCES IN LARGE MARGIN CLASSIFIERS*, pages 61–74. MIT Press.

Patrick Ruch, Julien Gobeill, Christian Lovis, and Antoine Geissbühler. 2008. Automatic medical encoding with snomed categories. In *BMC medical informatics and decision making*, volume 8, pages 1–8. BioMed Central.

Guergana K Savova, James J Masanz, Philip V Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C Kipper-Schuler, and Christopher G Chute. 2010. Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, 17(5):507–513.

Vincent A Traag, Paul Van Dooren, and Yurii Nesterov. 2011. Narrow scope for resolution-limit-free community detection. *Physical Review E*, 84(1):016114.

Vincent A Traag, Ludo Waltman, and Nees Jan Van Eck. 2019. From louvain to leiden: guaranteeing well-connected communities. *Scientific reports*, 9(1):1–12.

Jason W. Wei and Kai Zou. 2019. EDA: easy data augmentation techniques for boosting performance on text classification tasks. *CoRR*.