Taxonomy-Based Automation of Prior Approval using Clinical Guidelines

Saranya Krishnamoorthy, Ayush Singh
inQbator AI at eviCore Healthcare
Evernorth Health Services
firstname.lastname@evicore.com

Abstract

Performing prior authorization on patients in a medical facility is a time-consuming and challenging task for insurance companies. Automating the clinical decisions that lead to authorization can reduce the time that staff spend executing such procedures. To better facilitate such critical decision making, we present an automated approach to predict one of the challenging tasks in the process called *primary indicator* prediction, which is the outcome of this procedure. The proposed solution is to create a taxonomy to capture the main categories in primary indicators. Our approach involves an important step of selecting what is known as the "primary indicator" – one of the several heuristics based on clinical guidelines that are published and publicly available. A taxonomy-based PI classification system was created to help in the recognition of PIs from free text in electronic health records (EHRs). This taxonomy includes comprehensive explanations of each PI, as well as examples of free text that could be used to detect each PI. The major contribution of this work is to introduce a taxonomy created by three professional nurses with many years of experience. We experiment with several state-of-the-art supervised and unsupervised techniques with a focus on prior approval for spinal imaging. The results indicate that the proposed taxonomy is capable of increasing the performance of unsupervised approaches by up to 10 F1 points. Further, in the supervised setting, we achieve an F1 score of 0.61 using a conventional technique based on term frequency–inverse document frequency that outperforms other deep-learning approaches.

1 Introduction

Real-world applications in the Natural Language Processing (NLP) domain are known to perform better when the language models that support them are trained and fine-tuned on the domain in question (Gu et al., 2021; Rojas et al., 2022; Zhou et al., 2022; Naseem et al., 2022). One domain where this idea is applicable at a high level is the healthcare domain. Applications herein must adhere to the domain-specific vocabulary and guidelines. Prediction tasks require large amounts of sensitive data that contain information about patients and other details about the facilities that provide treatment. While the data sensitivity and protection challenges alone can be considered overwhelming due to the caveats of anonymization and privacy efforts, other atypical challenges based on knowledge and representation add to the complexity of NLP solutions in healthcare.

Knowledge from overworked staff, such as nurses and physicians, is critical to obtaining high-quality corpora to train NLP models. Due to the lack of time, medical personnel are often unwilling to participate in annotation tasks to transfer knowledge (Ishikawa, 2022; Fiałek, 2022; Aycock, 2022; Miley, 2022). Furthermore, when staff can participate in annotation tasks, facilities are usually unwilling to release annotations for public consumption, making their use by other healthcare systems extremely difficult. In this work, we present several experiments (unsupervised and supervised) using state-of-the-art (SOTA) deep-learning techniques and compare them to more conventional techniques like term frequency–inverse document frequency (TF-IDF). The experiments predict what is known as the "primary indicator" from a set of clinical guidelines for spinal imaging that are readily available on the Web[1]. The primary indicator is the first step of several for determining whether or not a patient should be approved for a spinal imaging procedure. Typically, indicators consist of findings

[1] Retrieved July 31, 2023, from https://www.nccn.org/guidelines/guidelines-with-evidence-blocks

Proceedings of Recent Advances in Natural Language Processing, pages 598–607
Varna, Sep 4–6, 2023
https://doi.org/10.26615/978-954-452-092-2_065

such as the presence of *pain*, *trauma*, and *fracture* when approval is required by a facility to perform a procedure.

Primary indicators of spine injuries are generally chosen by clinical personnel in a facility without automation using carefully prepared guidelines written by highly skilled physicians in the field. As a way of narrowing down the guidelines for language model prediction and facilitating future iterations of machine learning experiments, our work introduces a taxonomy available for public use annotated by three clinical professionals skilled in the area of nursing. Our experiments show that the use of taxonomy from skilled professionals can be used to increase performance for the real-world task at hand, especially in an *unsupervised* manner. The annotations created in this work are for use by the medical NLP community for investigative purposes and can be considered the main contribution therein.

Although we can achieve high F1 performance using transformer models (Devlin et al., 2018) on common corpora known as PubMed (Fiorini et al., 2018) and or MIMIC-III (Johnson et al., 2016). Due to this training procedure, these models are capable of performing well in biomedical corpora such as BC5CDR, I2b2, and others. However through this work we demonstrate on the contrary, that traditional models built on TF-IDF typically outperform deep learning models in terms of performance on unstructured corpus of insurance claims. We also contrast various fastText (Joulin et al., 2016) based models for unsupervised approaches with and without the added taxonomy.

The rest of the paper is organized as follows. First, we provide an overview of the limited existing work in this domain in section 2. Next, in section 3, we outline the problem that we aim to address. The construction of the taxonomy and annotation approaches is explained in Section 4.2. Subsequently, in Section 5, we discuss the approaches employed, along with the experimental details. Finally, we present a comprehensive analysis of the results in Section 6, and discuss future work in Section 7.

2 Related Work

Work that deals with real-world clinical data is sufficiently limited due to the prohibitive nature and sensitivity of facilities and patients. Most models and published work use some form of fine-tuning on models trained with corpora like the PubMed (Fiorini et al., 2018) and MIMIC-III (Johnson et al., 2016). However, the approaches presented in this section, while not comprehensive, cover SOTA approaches in the supervised and unsupervised clinical domain.

Supervised - various techniques such as self-supervised and contrastive learning are used by different studies. SapBERT (Liu et al., 2020), a self-supervised model, uses a transformer-based language model and a knowledge graph known as UMLS (Bodenreider, 2004) to classify entities of names. In their approach, they do not use clinical-based guidelines. In other work, they used masked-language modeling (MLM) called NeighBERT (Singh et al., 2022) that is capable to classify entities and link them using the UMLS as a guide. Our work does not use the UMLS – our intent is to provide a nurse-based taxonomy and several baseline approaches and to show the impact of the taxonomy without the complexities of finding entities. Our supervised approaches include two other commonly-used approaches known as BioBERT (Lee et al., 2020) and BlueBERT (Peng et al., 2019a). The BioBERT (Lee et al., 2020) model uses weights trained on a general domain from Wikipedia and the Google Books Corpus (Michel et al., 2011) and then pre-trains it using PubMed (Fiorini et al., 2018) abstracts. BlueBERT is similar to BioBERT with the additional inclusion of the MIMIC-III (Johnson et al., 2016) corpus in the fine-tuning procedure. In this work, we use both models and fine-tune them on our datasets. Our data comprise an unstructured corpus of insurance claims in the form of free text, which includes patient health records vital to making a decision of whether or not a claim should be approved.

Unsupervised work generally relies on methods of clustering along with the usage of external sources of knowledge like taxonomies or structured data such as UMLS. Target classes and input data are encoded using the same embeddings, and a distance measurement like cosine similarity is used to calculate the similarity between class representation and the input data. Embeddings can be created at the word, sentence, or document level. BioSentVec (Chen et al., 2019) and BioWordVec (Zhang et al., 2019) can both be used to generate embeddings. Some SOTA work uses BioWordVec (Amorim, 2022; Mao and Fung, 2020; El-Shimy et al., 2022) for both supervised and unsupervised tasks approaches. We use BioWordVec in our work to compare and contrast with other techniques such as BioSentVec (Chen et al., 2019). The results of other works that used BioWordVec suggest that this embedding performs well in unsupervised setting (Chen et al., 2019; Deka and Jurek-Loughrey, 2022; El-Shimy et al., 2022).

We found little work for the clinical domain that uses taxonomies together with embeddings. However, work from Kwon et al. (2022) is quite similar to ours because it uses BioSentVec (Chen et al., 2019) to create embeddings and has a classifier, albeit supervised, for named-entity recognition (NER). In their work, the task was based on entity finding, similar to NeighBERT (Singh et al., 2022) and others; here, we forego supervision outside of the annotations that are created. Other work (Lee et al., 2022) uses BioWordVec (Zhang et al., 2019) in a similar way with a supervised model. The majority of other related work that uses a taxonomy is based on a clustering technique such as Latent Dirichlet Allocation (LDA) (Blei et al., 2003). This form of clustering first clusters words from groups of documents and topics as a form of weak supervision, in which topics can be mapped to a taxonomy. Since we already have a taxonomy, LDA does not add any value to the data or to our approach, hence, we do not use LDA in this work.

3 Problem Statement

The overall objective is to mimic the behavior of clinicians in the prior authorization process. As an initial step, this research aims to address the aforementioned challenges and develop a multi-class classification approach that can accurately predict one of the 34 primary indicators from electronic health records. Ultimately, our goal is to improve the efficiency and effectiveness of information retrieval and knowledge management in the healthcare domain.

4 Taxonomy, Data Acquisition, and Annotation Methods

These sections elaborate on the annotation process and the taxonomy of the corpora provided.

4.1 Taxonomy and Annotation Methods

The main contribution of this work is to create a taxonomy comprising of a short description of each PI from the clinical guidelines. Overall three subject matter experts participated in the annotation task. All annotators had access to the publicly available guidelines[2] and were asked to produce two paragraphs of explanation related to primary indicators for spinal imaging assigned according to their experience explained in Section 4.2. The explanatory paragraphs were carefully reviewed to avoid the inclusion of sensitive data. For writing the paragraphs, the annotators were asked to use previous patient reports, documents, and other clinical material that would be used to determine a primary indicator. These documents are not publicly available – the taxonomy consists of the annotator's description summaries from the document structure and the taxonomy descriptions. We do not perform and discuss any inner-annotator agreement (IAA) due to the task being text generation, and it is not easy to measure a metric that shows a fair and unbiased IAA. However, as a litmus test, annotators were asked to work on the same 5 primary indicators (in blind tests).

To show the impact of the taxonomy we design a number of experiments which we explain in section 5 and results are discussed in section 6. To be able to share the taxonomy we obfuscated the text and redact any personal information such as gender, age, and individual stories. The changes are minor and will not affect the reproducibility of this work. We replaced the gender pronoun *he/she* with *they, the patient, patient* when applicable. Statements like, e.g., *65 year old women* change to *the patient between 63-68 year old.* Individual stories which are on average 10 tokens are taken out from the description. There are only a handful of individual stories which are irrelevant to their corresponding taxonomy. Finally, geographic and temporal information is replaced with `[LOCATION]` and `[DATE]` tags.

4.2 Annotator Details

The first annotator (Annotator 1) is a nurse with 14 years of clinical experience, 3 of which have been spent working in a clinical role for a private company. The annotator has less than 1 year of annotation experience. The annotator's main clinical experience is in cardio-pulmonary and emergency room documentation. Additionally, the annotator has worked on clinical surveys based on the clinical guidelines used for experimentation. The annotator

[2] Retrieved July 31, 2023, from https://www.evicore.com/provider/clinical-guidelines

Corpus	Train	Dev	Test
Number of documents	190655	23832	23832
Number of tokens	328M	41M	41M
Number of sentences	13.2M	1.6M	1.6M
Mean number of tokens per document	1723	1728	1735
Mean number of sentences per document	69	70	68

Table 1: Statistics for the training, development, and test corpus used in experiments.

is well-versed in guideline reading and writing for healthcare systems and has completed several tasks for the company used for experimentation.

The second annotator (Annotator 2) is a nurse with 13 years of clinical experience, 6 of which have been spent in a private enterprise clinical role. The annotator has approximately 9 years of experience in healthcare annotation. The annotator's main clinical experience is in maternal, cancer, neonatal, ICU, and electronic health record (EHR) documentation. Additionally, the annotator has peer-reviewed clinical surveys for major systems. The annotator is well-versed in guideline reading and writing for healthcare systems and has completed several tasks for the company used for experimentation.

The third annotator (Annotator 3) is a nurse with 28 years of clinical experience, 3 of which have been spent working in a private enterprise clinical role. The annotator has less than 1 year of annotation experience. The annotator's main clinical experience is in labor and delivery, emergency department, vascular access, OB / GYN and gastroenterology. Additionally, the annotator has worked on clinical surveys based on the clinical guidelines used for experimentation. The annotator is well-versed in guideline reading and writing for healthcare systems and has completed several tasks for the company used for experimentation.

4.3 Corpus Collection

We use a corpus collected from several real-world prior authorization data sources. The corpus itself comprises of patient notes in the form of unstructured free text found in the electronic health record of the patient. The clinical staff uses the same free text when they try to ascertain which *primary indicator* the patient exhibits. Although the corpus could not be publicly released due to PHI restrictions, the taxonomy produced by nurses is available[3]. Furthermore, vital corpus statistics are reported in Table 1 where the corpus is split into training, development, and test sets.

5 Modeling "primary indicator" (PI)

Experiments are broken down into several tasks related to SOTA in the field covered in Section 2. Specifically, we separate the settings into two types: *Supervised* and *Unsupervised* to show the benefit of the taxonomy while also applying the latest techniques to solve the real-world problems at hand. We first set baselines of how far supervised techniques can reach before moving on to showing the advantage of using our method on unsupervised techniques. The following two sections explain the supervised and unsupervised experiment settings. To evaluate our models, we use weighted F1 score in order to account for the high-class imbalance present in corpus (see Appendix 3 for details). All the hyperparameters for the aforementioned approaches are detailed in the Appendix Table 4.

5.1 Supervised

There are two baseline models used during experimentation. Both baseline models use a random-forest classifier (RFC) (Breiman, 2001) for classification on output from two word representation algorithms: a TF-IDF and bag-of-words (BOW) model. These are selected because oftentimes clinical text would have critical keywords required for reasoning and semantic representation might not be needed. A hyper-parameter grid search is used to find the optimum hyper-parameters for the RFC and the best performing model for both models (TF-IDF and BOW) is reported for comparison.

For other approaches that do take semantics into account, we fine-tune a BioWordVec (Zhang et al., 2019) model on the training data to create token-based word embeddings. The embeddings are then

[3] Retrieved July 31, 2023, from https://github.com/inQbator-eviCore/clpt/taxonomy

used as input to an RFC (Breiman, 2001) trained to classify among the various 34 classes. Similarly, we experiment with sentence-level embeddings using BioSentVec(Chen et al., 2019) by extracting embeddings and using them as input to a convolutional neural network (CNN) model. Additionally, we experiment with BioBERT (Lee et al., 2020) model pre-trained on PubMed (Fiorini et al., 2018) text. We also experiment with BlueBert (Peng et al., 2019b) which is trained on both PubMed and the MIMIC-III (Johnson et al., 2016) dataset. Fine-tuning of both BERT models is performed using the training data discussed in Section 4.3.

5.2 Unsupervised

We used two-sentence embedding models to perform unsupervised classification in two experimental settings: *with and without taxonomy*. We used the introduced taxonomy from nurse annotators for the *with taxonomy* experiment and we use the text from the clinical guidelines alone in a “cut and paste” manner for the *without taxonomy* experiment.

In order to measure the distance between the patient report text and the introduced taxonomy we split both the input text and the annotated text into sentences. We use the cosine similarity distance, a vector space measurement used to find semantic similarity in the past (Rahutomo et al., 2012), to determine which target sentences (or labels) are most similar to the input sentences in the document. For the target sentences, we combine the sentence-based vectors and calculate the mean to make sure the dimensions of the resulting vector stay the same. An exhaustive search is performed for each sentence in the input text and the most similar sentences are used for classification. We hypothesize that this approach will lead to an evidence-based approach in future work where the sentence most similar to the sentence would be presented as evidence. Sentence embeddings are created using BioSentVec (Chen et al., 2019) and compared to fastText-based (Joulin et al., 2016) Sent2Vec (Moghadasi and Zhuang, 2020). Both are trained using the training data, and all parameters are defined in Table 4.

6 Results

In this section, we present our experimental results for both the supervised and unsupervised approaches in Table 2. The use of a taxonomy for supervised experiments is saved for future work. Nonetheless, we demonstrate the effectiveness of the taxonomy introduced with a comparison that uses cosine similarity as the measurement of the distance between the input sentence and the target primary indicator description (created by the nursing annotators).

Supervised			
	Precision	Recall	Weighted F1
BOW + RFC	0.59	0.62	0.52
TFIDF + RFC	0.66	0.66	0.61
BioWordVec + RFC	0.57	0.56	0.49
BioSentVec + CNN	0.43	0.58	0.48
BioBERT	0.49	0.66	0.56
BlueBERT	0.53	0.62	0.57
Unsupervised			
FastText	0.54	0.02	0.04
BioSentVec	0.37	0.02	0.03
FastText + taxonomy	0.43	0.07	**0.12**
BioSentVec + taxonomy	0.38	0.08	**0.13**

Table 2: Comparison of supervised and unsupervised approaches with and without the nurse’s taxonomy contribution. The unsupervised approaches see a significant boost in performance after the addition of taxonomy data.

Under supervision, the TF-IDF model outperforms other deep-learning models based on BERT (Devlin et al., 2018) and fastText (Joulin et al., 2016). This is due to the fact that other approaches like BioWordVec (Zhang et al., 2019) and BioSentVec (Chen et al., 2019) on average have a 48 percent out-of-vocabulary (OOV) word detriment. This forces pre-trained word-embedding models to perform poorly when the words are not available. While models based on the BERT (Devlin et al., 2018) architecture are typically known to outperform conventional models such as TF-IDF, the limitation of 512-word tokens for these experiments degrades the resulting performance. In our corpus, the documents are generally about three times larger than the 512-word-token limit (an average of 1700 tokens). In this real-world setting, the adaptation of the baseline NLP models was necessary along with the experimentation of taxonomy to better understand the value of knowledge representation for the task.

As shown in Figure 1, the unsupervised approaches including both Sent2Vec (Moghadasi and Zhuang, 2020) and fastText (Joulin et al., 2016) show that the use of the introduced taxonomy increases the performance considerably when compared to a sim-

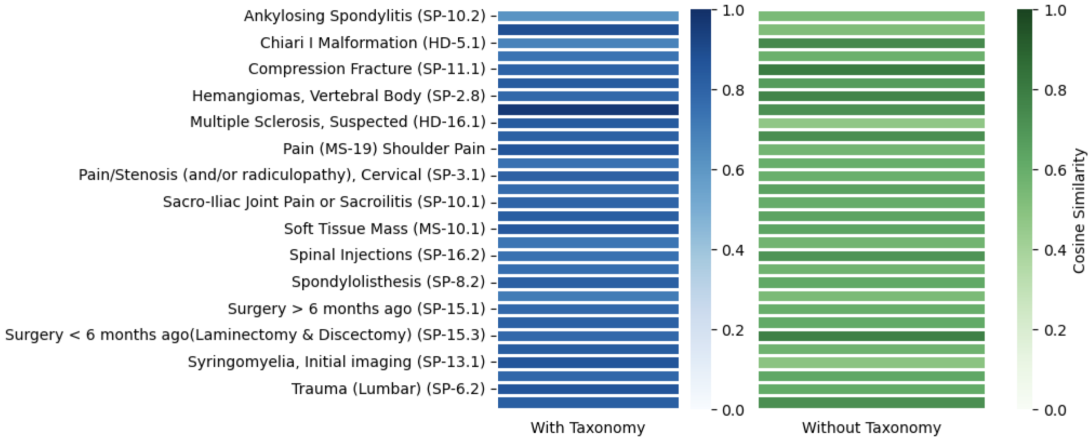

Figure 1: Averaged Cosine similarity measurements for all primary indicators showing signals received from taxonomy vs "cut and paste" from clinical guidelines e.g. Primary indicators like Multiple Sclerosis, Suspected HD-16.1 shows stronger signal when using our taxonomy.

ple "cut and paste" approach directly from the clinical guidelines. We also note that when the fastText (Joulin et al., 2016) model is trained on our training data, it outperforms other off-the-shelf approaches like BioSentVec (Chen et al., 2019) when using the introduced taxonomy. We believe that the underperformance is due to both the domain and the lack of vocabulary (covering only nearly 50% of the vocabulary in the test set).

The nursing annotations are somewhat more descriptive for Annotators 2 and 3. We believe that this is due to the domain knowledge. However, in some cases, Annotator 1 described more specific cases. Another note that we should present – annotators did not annotate for 3 classes [HD-16.1, SP-2.2, SP-2.8]. This was due to the fact that those primary indicators were irrelevant and are not currently used in the clinical guidelines. In our experiments, we excluded those primary indicators from all sets. Annotators also indicated that the *Inflammatory Spondylitis* primary indicator is nearly the same as *Ankylosing Spondylitis* class. In that case, we updated both class labels as one only.

7 Conclusion and Future Work

We introduce a novel corpus-based taxonomy from a real-world clinical setting. This taxonomy is created from publicly available guidelines and used as a corpus of instrumentation in an unsupervised setting. The corpus itself, created by three nursing annotators with several years of experience, illustrates how domain knowledge can increase the performance of the spinal imaging primary indicator in a set of clinical guidelines (also public).

Experiments in the supervised setting show that we are able to achieve decent F1 results with state-of-the-art techniques based on deep learning. Our next steps are to include the taxonomy in the supervised setting in hopes of achieving F1 scores of at least 80% which will make this approach viable to use in a real-world setting. Additionally, we intend to create a classifier that is capable of processing further indications from the clinical guidelines.

Acknowledgements

We would like to thank Deb Fox, Nicole Van Loan, Rhonda Laxton and Katherine Riggs for their crucial support in creating the taxonomy. We also thank Shabhnam Tafreshi, Shubham Vatsal and Navpreet Singh for their crucial feedback that helped in shaping the research direction. We further thank the support of the enterprise in funding the research.

Ethics Statement

The authors of this article have set out to purposely create a worthwhile contribution to the scientific community by creating a taxonomy with the help of actual nurses in the clinical domain. We provide the taxonomy and the code in a framework (Krishnamoorthy et al., 2022)[4] and request the community to please report to us via email for any further advancements.

[4] Retrieved July 31, 2023, from https://github.com/inQbator-eviCore/clpt

References

Sofia Pessoa de Amorim. 2022. *Evaluating Pre-trained Word Embeddings in domain specific Ontology Matching*. Ph.D. thesis.

Ryan Aycock. 2022. Overworked nurses need relief. *Emergency Medicine News*, 44(2):7–8.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Olivier Bodenreider. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270.

Leo Breiman. 2001. Random forests. *Machine learning*, 45:5–32.

Qingyu Chen, Yifan Peng, and Zhiyong Lu. 2019. Biosentvec: creating sentence embeddings for biomedical texts. In *2019 IEEE International Conference on Healthcare Informatics (ICHI)*, pages 1–5. IEEE.

Pritam Deka and Anna Jurek-Loughrey. 2022. Evidence extraction to validate medical claims in fake news detection. In *Health Information Science: 11th International Conference, HIS 2022, Virtual Event, October 28–30, 2022, Proceedings*, pages 3–15. Springer.

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*.

Heba El-Shimy, Hind Zantout, and Hani Ragab Hassen. 2022. Assessment of pharmaceutical patent novelty with siamese neural networks. In *Artificial Neural Networks in Pattern Recognition: 10th IAPR TC3 Workshop, ANNPR 2022, Dubai, United Arab Emirates, November 24–26, 2022, Proceedings*, pages 140–155. Springer.

Bartosz Fiałek. 2022. On the verge of poland's fifth wave of covid-19, healthcare staff are overworked and disenchanted.

Nicolas Fiorini, Robert Leaman, David J Lipman, and Zhiyong Lu. 2018. How user intelligence is improving pubmed. *Nature biotechnology*, 36(10):937–945.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23.

Masatoshi Ishikawa. 2022. Overwork among resident physicians: national questionnaire survey results. *BMC Medical Education*, 22(1):729.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9.

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016. Fasttext. zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.

Saranya Krishnamoorthy, Yanyi Jiang, William Buchanan, Ayush Singh, and John Ortega. 2022. CLPT: A universal annotation scheme and toolkit for clinical language processing. In *Proceedings of the 4th Clinical Natural Language Processing Workshop*, pages 1–9, Seattle, WA. Association for Computational Linguistics.

Sunjae Kwon, Zhichao Yang, and Hong Yu. 2022. An automatic soap classification system using weakly supervision and transfer learning. *arXiv preprint arXiv:2211.14539*.

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.

Sang-Woo Lee, Nam Kim, Jung-Hyok Kwon, Hyung Do Choi, Sol-Bee Lee, and Eui-Jik Kim. 2022. Comparative study of word embeddings for classification of scientific article on human health risk of electromagnetic fields. In *2022 IEEE 11th Global Conference on Consumer Electronics (GCCE)*, pages 391–392. IEEE.

Fangyu Liu, Ehsan Shareghi, Zaiqiao Meng, Marco Basaldella, and Nigel Collier. 2020. Self-alignment pretraining for biomedical entity representations. *arXiv preprint arXiv:2010.11784*.

Yuqing Mao and Kin Wah Fung. 2020. Use of word and graph embedding to measure semantic relatedness between unified medical language system concepts. *Journal of the American Medical Informatics Association*, 27(10):1538–1546.

Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K Gray, Google Books Team, Joseph P Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, et al. 2011. Quantitative analysis of culture using millions of digitized books. *science*, 331(6014):176–182.

Viv Miley. 2022. Overworked nurses protest conditions. *Green Left Weekly*, (1332):4.

Mahdi Naser Moghadasi and Yu Zhuang. 2020. Sent2vec: A new sentence embedding representation with sentimental semantic. In *2020 IEEE International Conference on Big Data (Big Data)*, pages 4672–4680. IEEE.

Usman Naseem, Byoung Chan Lee, Matloob Khushi, Jinman Kim, and Adam G Dunn. 2022. Benchmarking for public health surveillance tasks on social media with a domain-specific pretrained language model. *arXiv preprint arXiv:2204.04521*.

Yifan Peng, Shankai Yan, and Zhiyong Lu. 2019a. Transfer learning in biomedical natural language processing: an evaluation of bert and elmo on ten benchmarking datasets. *arXiv preprint arXiv:1906.05474*.

Yifan Peng, Shankai Yan, and Zhiyong Lu. 2019b. Transfer learning in biomedical natural language processing: An evaluation of BERT and elmo on ten benchmarking datasets. *CoRR*, abs/1906.05474.

Faisal Rahutomo, Teruaki Kitasuka, and Masayoshi Aritsugi. 2012. Semantic cosine similarity. In *The 7th international student conference on advanced science and technology ICAST*, volume 4, page 1.

Matías Rojas, Jocelyn Dunstan, and Fabián Villena. 2022. Clinical flair: a pre-trained language model for spanish clinical natural language processing. In *Proceedings of the 4th Clinical Natural Language Processing Workshop*, pages 87–92.

Ayush Singh, Saranya Krishnamoorthy, and John Ortega. 2022. Neighbert: Medical entity linking using relation induced dense retrieval.

Y Zhang, Q Chen, Z Yang, HF Lin, and ZY Lu. 2019. Biowordvec, improving biomedical word embeddings with subword information and mesh. sci data 6: 52.

Sicheng Zhou, Nan Wang, Liwei Wang, Hongfang Liu, and Rui Zhang. 2022. Cancerbert: a cancer domain-specific language model for extracting breast cancer phenotypes from electronic health records. *Journal of the American Medical Informatics Association*, 29(7):1208–1216.

A Class Imbalance

Table 3 shows the high class-imbalance ratio of almost 1000 times between the majority and minority class present in the corpus. It can observed that the top-5 indicator codes make up almost 90% of the volume in the corpus.

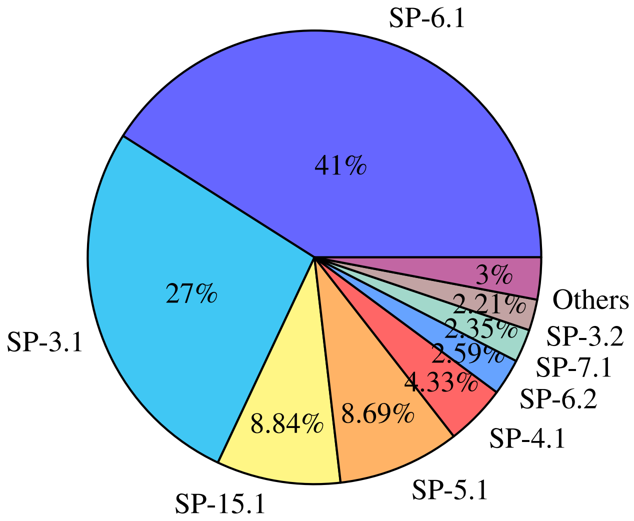

Figure 2: Pie chart showing class distribution in the corpus. The rest of 26 classes only cover about 3% of the volume.

Table 3: A table showcasing extreme class imbalance present in the dataset

Primary Indication	**Count**
Lower Extremity Pain (with radiculopathy), with or without Low Back Pain (SP-6.1)	96706
Pain/Stenosis (and/or radiculopathy), Cervical (SP-3.1)	64421
Surgery greater than 6 months ago (SP-15.1)	21079
Pain (without radiculopathy), Lumbar (SP-5.1)	20720
Pain/Stenosis (and/or radiculopathy), Thoracic (SP-4.1)	10319
Trauma (Lumbar) (SP-6.2)	6162
Myelopathy (SP-7.1)	5598
Trauma (Cervical) (SP-3.2)	5270
Compression Fracture (SP-11.1)	1212
Spinal Stenosis, Lumbar (SP-9.1)	1131
Trauma (Thoracic) (SP-4.2)	958
Surgery less than 6 months ago (Fusion) (SP-15.3)	581
Spinal Lesion, Other (SP-2.8)	574
Surgery less than 6 months ago (Laminectomy and Discectomy) (SP-15.3)	517
Spondylolisthesis (SP-8.2)	470
Multiple Sclerosis, Known (HD-16.1)	441
Multiple Sclerosis, Suspected (HD-16.1)	423
Scoliosis or Kyphosis (SP-14.1)	338
Spinal Cord Stimulator Placement/Removal (SP-16.3)	312
Syringomyelia, Initial imaging (SP-13.1)	154
Ankylosing Spondylitis (SP-6.2)	136
Soft Tissue Mass (MS-10.1)	91
Spondylolysis (SP-8.1)	88
Syringomyelia, Follow up imaging (SP-13.2)	85
Spinal Injections (SP-16.2)	80
Ankylosing Spondylitis (SP-10.2)	71

Continued on next page

Table 3: A table showcasing extreme class imbalance present in the dataset (Continued)

Hemangiomas, Vertebral Body (SP-2.8)	64
Chiari I Malformation (HD-5.1)	61
Inflammatory Spondylitis (SP-10.2)	53
Chronic/Stable Spine Pain (SP-1.0)	50
Positional MRI (SP-2.2)	50
Headache (HD-11)	49
Pain (MS-19)	28
Sacro-Iliac Joint Pain or Sacroilitis (SP-10.1)	27

B Hyperparameters

Approaches	**Parameters**
TF-IDF	We use bi-grams along with L2 regularization and maximum document frequency set to 0.75 and minimum document frequency of 0.10.
BOW	We use bi-grams with a maximum document frequency of 0.80 and minimum document frequency of 0.10.
RFC	Baseline experiments are run with a random forest classifier (RFC) and bootstrapping. Split quality of the classifier is measured using entropy and tree depth is set to 85 along with a tree count of 90.
BioWordVec	The BioWordVec (Zhang et al., 2019) classifier is trained using FastText (Joulin et al., 2016) using 200 dimensions and a six-gram word size. Learning rate is set to 0.001. A window size of 30 is used along with 10 negative sample size.
BioSentVec	The BioSentVec (Chen et al., 2019) classifier is trained using the Sent2Vec (Moghadasi and Zhuang, 2020) algorithm. We use a 700 dimension matrix size along with a bi-gram representation. Dropout is set to 0.001 and sampling of 10 negative samples combined with a window size of 30.
CNN	Both BiowordVec and BioSentVec (Chen et al., 2019) classifiers use a convolutional neural network (CNN). The CNN used three layers and filter sizes ranging from 3-5 and 100 filters for each layer. Optimization is based on the Adam's optimization using a learning rate of .0001. Both classifiers are trained for 10 epochs with a dropout set to 0.5. Fine-tuned using other BERT models are fine-tuned using with 50 epochs and early stopping. The learning rate is set to 0.001 starting with 0.1 and reducing by factors of .10 whenever loss plateaus consecutively for three epochs.
FastText Sent2Vec	A Sent2Vec (Moghadasi and Zhuang, 2020) model is using for training. A matrix size of 700 dimensions is applied along with a bi-gram word size. Dropout is set to 0.001 with negative sampling set to 10 and the use of a window size of 30.

Table 4: Hyper-parameters used for the supervised and unsupervised models.