Unstructured Data Abstraction utilizing Selective Prediction-Oriented Neural Networks in Healthcare Settings

Stanford CS224N Custom Project

Emily Chen
Department of Computer Science
Stanford University
emilyc02@stanford.edu

Nathan Mohit
Department of Computer Science
Stanford University
mohitn@stanford.edu

Nicole Tong
Department of Computer Science
Stanford University
nwtong@stanford.edu

Abstract

This project focuses on revolutionizing the extraction of valuable insights from unstructured clinical data in healthcare. Many existing machine learning and deep learning models are supervised and thus require labeled data for training. Annotating unstructured clinical data can be challenging because of the task's cognitive complexity and the variability in the data quality. We aim to leverage a multi-layer perceptron (MLP) to parse through unstructured clinical narratives and output the statement of diagnosis, used by past work such as Conneau et al. (2020). Such tasks include natural language processing techniques such as text classification, information extraction, and selective prediction. This project to contribute to this immense potential for improving medical decision-making and patient care.

1 Introduction

In the healthcare system, electronic health records hold mountains of information and data in a combination of two formats - structured and unstructured data. Structured data presents as information specifically selected by clinicians from predetermined options, often used for medication prescriptions, bare biological data (gender, height, age), and other fields with intuitive choices. Unstructured data takes the form of clinician's prose or bare notes- essentially anything that clinicians would record as memorable/significant that can't easily be fit into a predetermined field/that the clinician themselves choose to record in an unstructured format (likely due to efficiency). These unstructured notes often hold extremely valuable information such as the observations and treatment plans devised on a patient's first visit, patient history, response to treatments, symptom progression, and disease recurrence.

Many professionals in the medical field choose to record their observations and thoughts in an unstructured format due to ease and efficiency as opposed to structured formats that may require much more effort than necessary to treat a patient. The result from this is personalized unstructured data particular to the author that is not easily translatable, nor easily analyzed for further downstream analysis. Manual abstraction of data from these health records is logistically infeasible, as it is timely, costly, and overall viewed as purposelessly infinite as records are altered and made at an exponential rate. However, the insights to be gained for healthcare in this unstructured data is indisputably invaluable and can open the door to innovation in almost every division of healthcare.

In the realm of healthcare, handling unstructured clinical data has always been a computationally expensive and laborious endeavor. Previous approaches have used manual abstraction and structured proxy variables, but these methods may be time-consuming, unscalable, and inaccurate.

Our goal to navigate unstructured clinical data to provide a diagnose with a degree of uncertainty using a multi-layer perceptron would impact the lives of both doctors and patients in 1) delivering results and 2) making diagnoses in a faster and more accurate way. Instead of classifying the unstructured data as a solidified diagnosis, there would be an uncertainty range in which the case would be flagged. This decision comes in line with the Hippocratic Oath of doing no harm, which we also hope to implement in our project (to classify as accurately possible, but also recognize uncertainty).

In the next section, we will specify previous approaches to selective prediction in unstructured clinical data and how our project extends upon base techniques.

2 Related Work

In recent years, ongoing research in the intersection of healthcare and NLP is the the utilization of selective prediction within the model-assisted abstraction techniques currently being utilized. As manual clinical abstraction has been rendered suboptimal, the current research endeavors in the biomedical informatics space as well as the NLP space is to leverage model-assisted abstraction wherein machine learning applications prevail.

Li et al. (2022) discusses various neural methods and techniques applied to unstructured data abstraction, especially within electronic health records. The neural methods highlighted contemporary deep learning models like convolutional neural networks (CNNs) and recurrent neural networks (RNNs), for subtasks including medical text classification, segmentation, word sense disambiguation, and specialized tasks like medical coding, outcome prediction, and de-identification. The paper also mentions the use of pre-trained models such as BERT and BioBERT for classification tasks, showcasing the evolution from rule-based and feature-driven methods to deep learning and embedding-based approaches for handling the complexity of medical text data.

These machine learning models however do not hold high enough efficiency or accuracy when compared to manual abstraction, leading to the application of utilizing models simply to flag high confidence records. However, in a model that holds bias for these high confidence records, many other records may be missed but essential towards comprehensive and equal insights into the healthcare data being examined. Furthermore, in a healthcare space/context, high accuracy is especially crucial to a working model, as an incorrect classification can result in a direct negative effect to a real patient, meaning that it is often seen as better for no classification to be given to a data point rather than an uncertain one.

The solution that Conneau et al. (2020) chose to consequentially develop would incorporate the advantages of machine learning models with manual abstraction. The aim would herein be to train ML models to a high accuracy, but flag records that had a low confidence level to be reviewed by manual abstractors - ideally eliminating the risk of incorrect classifications. Following the feedback from manual abstraction, the results would be fed back into the model and lead to an ultimately more confident model in the end. The use of this theoretic utility-based thresholding would lead to improved accuracy and a more efficient workflow for abstraction overall.

However, most of the relevant work surrounding selective prediction has focused on base techniques, namely logistic regression, random forest, and support vector machine models, rather than neural networks. Other limitations in this space include only being tested on clinical oncology documents (pathology reports) and considering binary variables (diagnosis and procedures), meaning that further testing must be done to evaluate if the model can abstract non-binary variables in different contextual diseases.

Our project seeks to extend upon the selective prediction method with more advanced techniques using a multi-layer perceptron, a type of neural network.

3 Approach

Our approach follows advances by a multi-layer perceptron for the purpose of prediction under a supervised protocol. A MLP is a feed-forward neural artificial network with an input layer, one or more hidden layers, and one prediction layer at the top, for classification. For the multi-layer perceptron architecture, we first use input layer as sequences of word embeddings. The input layer represents the raw text data. Each input neuron corresponds to a feature in the text, such as word embeddings, TF-IDF scores, or other numerical representations derived from the text. Then, we apply multiple layers to capture hierarchical features in the clinical data. Each hidden layer consists of multiple neurons, using a Rectified Linear Unit (ReLU) activation function. Our output layer corresponded to a particular disease (in our case, depression).

As mentioned in Gangavarapu et al. (2020), the first layer takes $\mu^{(m)}$ with p' clinical terms as the input and uses the output of each layer as the input to the following layer. Thus, we can represent the transformation as the following: $O^{(l)} \rightarrow W^{(l+1)}O^{(l)} + b^{(l+1)} \rightarrow g(W^{(l+1)}O^{(l)} + b^{(l+1)}) \rightarrow O^{(l+1)}$

where l is the layer with output$O^{(l)}$, weights $W^{(l+1)}$, and biases $b^{(l+1)}$.

In order to train, MLP uses backpropagation, which is used to calculate the gradient of the loss function to update weights, which aids the MLP to learn the internal representations, allowing it to learn any arbitrary mappings within the network Gangavarapu et al. (2020). In the case of multi-label classification, while the forward pass remains the same, the backpropagation algorithm uses a global error function that addresses the dependencies between the class labels.

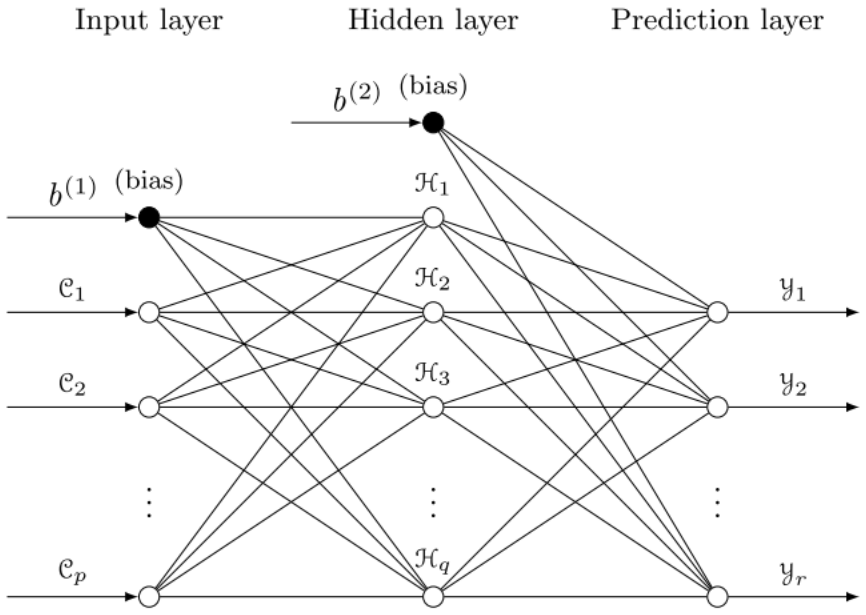
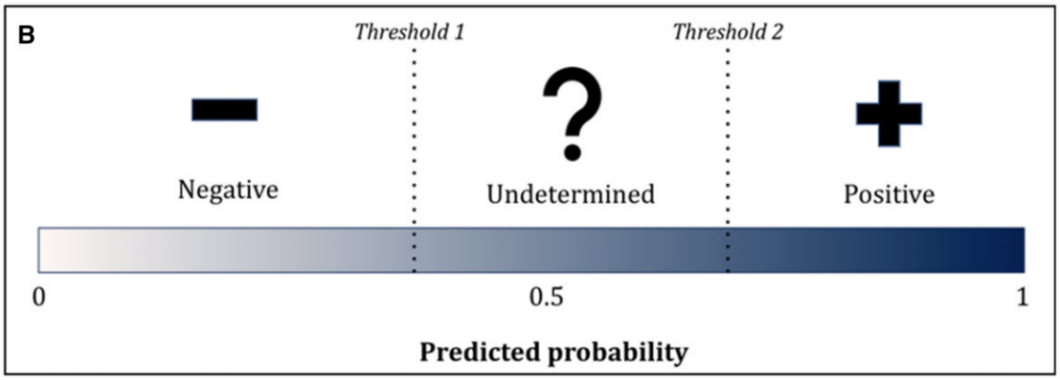


Fig. 6. Multi-label classification neural network model with p input clinical terms ($\mathcal{C}_i$s), a hidden layer with q nodes ($\mathcal{H}_i$s), and r possible ICD-9 code groups ($\mathcal{Y}_i$s).

This image from Gangavarapu et al. (2020) shows a one hidden layer feed-forward MLP network for multi-label classification. Next, we will detail our approach to include selective prediction in our MLP.



The above figure is a visual of the thresholds to determine whether a diagnosis is undetermined. Two thresholds (Threshold 1 and Threshold 2) are applied to a continuous predicted probability to yield 3 possible classifications: negative (predicted probability < Threshold 1), positive (predicted probability > Threshold 2), and undetermined (Threshold 1 predicted probability Threshold 2).

$$\text{Total cost} = \frac{\begin{array}{l}(\text{cost of FP})(\#\text{ of FP}) + (\text{cost of FN})(\#\text{ of FN}) \\ +(\text{cost of undetermined})(\#\text{ of undetermined})\end{array}}{\text{Total \# of observations}}, \quad (1)$$

Thresholds are selected based on real world utilities so as to minimize the total cost (Akshay Swaminathan (2024)).

3.1) Baselines We implemented two baselines. For the first baseline, we implemented a Support Vector Machine (SVM) algorithm, a supervised machine learning algorithm commonly used for classification and regression tasks. We did so by transforming the tokenized text into numerical features using Term Frequency-Inverse Document Frequency (TF-IDF) vectorization. This represents the importance of each word in the document relative to the entire dataset. The second baseline involved a simple neural network from our milestone that served as a normal classifier, as opposed to a selective classifier. Data to compare between the selective classification and normal classification was not utilized as the data developed by the simple neural network would not be an appropriate comparison to the larger neural network utilized in the selective classifier. Redoing the normal classification under these new parameters would be necessary for any complete analysis of the baselines to the normal neural network.

3.2) Code The initial starter code was developed off of a paper that analyzed different NLP classifiers on unstructured medical data from the Leiden University Medical Centre outpatient rhemuatology dataset. The paper preprocessed the dutch dataset and trained 7 general classifiers, importing a class/pipeline they created for the purpose of training and evaluating multiple models on the same dataset. While this starter code was extensive and a great start for our milestone in terms of creating models easily, the pipeline was not especially useful when it came to training the larger, selective neural network, as parameters were not easily changeable and could not be altered for the selective classification purposes. Kept code from this starter code mostly laid in the methodology of the data preprocessing, but this code was significantly altered for the structure and limits of our datasets. Additional code that was synthesized from new was the neural network design, training, and evaluation. In particular, the MLP Classification code that allowed for the selective prediction was especially meaningful to the goals of the project.

4 Experiments

- **Data**: The dataset utilized was the MIMIC-IV-Note, a collection of free text notes from clinical data systems, collected by MIT from Beth Israel Deaconess Medical Center's Metavision Electronic Medical Records system. The free text notes are designated discharge notes, created at the time of discharge for a patient and intended to hold a comprehensive overview of presentation, tests, treatments, and outcomes for every admittance to the hospital. These notes can be generated via the assembly of every personnel comment made in the patient's chart during the visit, but are more often free-text narrative summaries created by the discharging physician. Because of the potential differentiating diagnoses and data from repeated patients, data files were limited to each individual hospital admittance and discharge, ensuring that there was only one diagnosis made for the hospital visit and eliminating any potential crossover from patients at different points in time. The dataset was scrubbed of any patient identifying information, with each patient holding a specific subject ID and additionally each individual admittance holding its own ID as well. The MIMIC-IV Core dataset was utilized as well, specifically for its records of the patients' diagnoses and their respective diagnostic codes.

  The patient files that were chosen to make up the subset to evaluate the model on were based on their diagnoses - with a positive label attributed to patients diagnosed with ICD Code F329 for Major depressive disorder, single episode, unspecified. While any other diagnosis could have been utilized for a negative label, patients diagnoses with ICD Code F419 for Anxiety disorder, unspecified, were specifically chosen to make up the evaluation sets, as the similarity in lexicons utilized for both diagnoses and their respective treatments would likely train a stronger, more discriminate model, as opposed to alternative diagnoses which could have a simpler level of differing lexicons. In addition, as these two diagnoses could overlap in many patients and/or admittances, admittances and their textual notes were only included if a diagnosis for both diseases was not present in that specific admittance. In doing so, the evaluative dataset included hospital admittances with diagnoses for either depression exclusively or anxiety exclusively, with 1000 cases randomly being chosen of each, totaling to 2000 cases total–an equal and proportionate level of exposure to each diagnosis and free text file. Each text file was additionally preprocessed into word stems and had english stop words removed, as determined by the nltk stop word python package. (Maarseveen T, 2020)

- **Evaluation method**: We evaluated our model with metrics commonly used in text classification and prediction tasks, such as F1 score, accuracy, precision, and recall.

  We also compared the F1 scores of the different classifiers. The F1 score, a measure of predictive performance, is particularly useful in scenarios where false positives and false negatives have a significant impact on the overall model performance.

  This metric is important given the medical context of our project. False positives (or misdiagnoses) can lead to unnecessary treatments or interventions, causing harm or discomfort to patients, while false negatives (missed diagnoses) can result in delayed or missed treatment, potentially leading to worsened health outcomes or even fatalities.

  By considering both precision and recall, the F1 score captures the trade-off between correctly identifying positive cases and minimizing misclassifications, making it particularly suitable for medical tasks where achieving high accuracy alone may not be sufficient.

- **Experimental details**: We used scikit-learn's Tfidf Vectorizer to transform the text data into numerical vectors in order for our model to understand and process the text. Term Frequency-Inverse Document Frequency (TF-IDF) is a statistical measure used to evaluate the importance of a word within a document in relation to a collection or corpus of documents. Then, we started with a default Multi-Level Perceptron neural network, as defined by scikit-learn. From there, we restructured the network to include 5 hidden layers that would help with recognizing additional features extracted from the texts. The MLPClassifier used a softmax function for the final output. We conducted a 5-fold cross-validation and calculated the average of the out-of-sample accuracies to assess the performance of our model on new, previously unseen data.

  We implemented the selective prediction, taking thresholds from Akshay Swaminathan (2024), specifically the thresholds calculated for depression. If the predicted probability

of having a depression diagnosis for a patient was above 0.9, it was given a label of 1, for having a depression diagnosis. If the predicted probability was between 0.9 and 0.2, then we labeled those as unsure, in alignment with our goal of following the Hippocratic Oath of doing no harm. These files could then undergo manual labeling, to ensure patients are not harmed through false positives or false negatives. Finally, if the predicted probability was under 0.2, they were given a label of 0, for not having a depression diagnosis.

- **Results**: Our model achieved an accuracy of 0.74 before the selective prediction, and 0.81 with the selective prediction, in which the patients we labeled "undetermined" are not considered. The model achieved an F1 score of 0.83 with selective prediction.

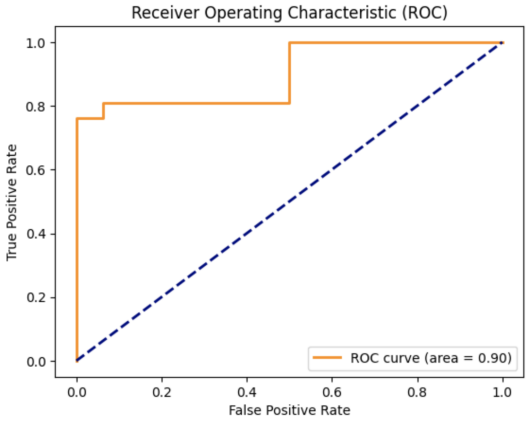


Figure 1: Receiver Operating Character graph to illustrate the trade-off between the true positive rate (TPR) and the false positive rate (FPR) across different threshold values

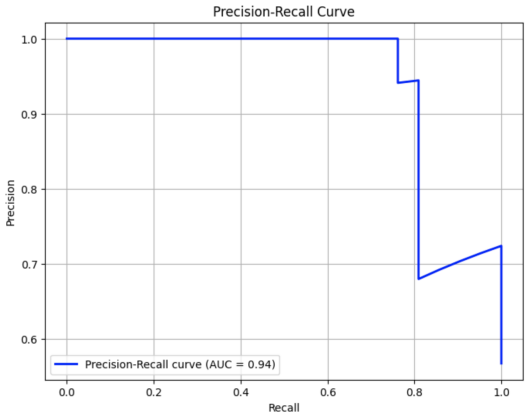


Figure 2: Precision Recall Curve. Precision measures the proportion of true positive predictions among all positive predictions made by the classifier. It is calculated as TP / (TP + FP), where TP is the number of true positives and FP is the number of false positives. Recall, also known as sensitivity or true positive rate (TPR), measures the proportion of actual positive samples that are correctly identified as positive by the classifier. It is calculated as TP / (TP + FN), where FN is the number of false negatives.

5 Analysis

Our model shows promise, but there is room for improvement in decreasing amounts of false negatives and false positives before and after the addition of the selective prediction functionality.

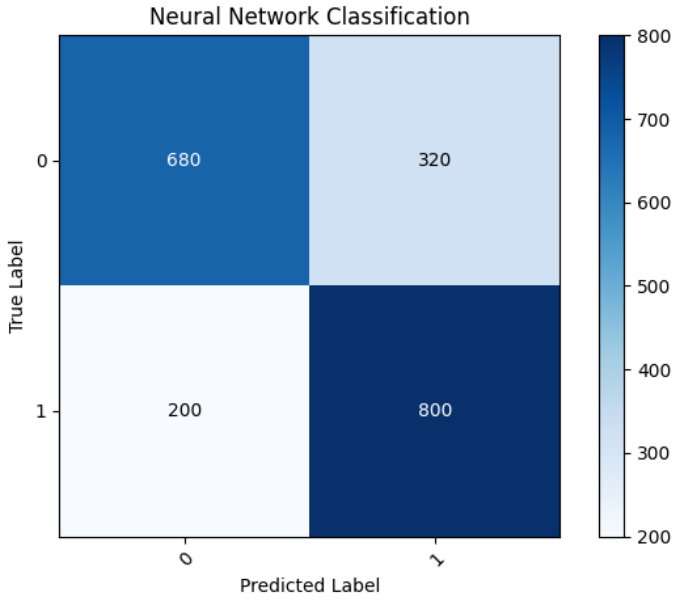


Figure 3: Confusion Matrix for Neural Network. The neural network on it's own was able to classify a high number of true positives and true negatives, with about 25% of the data being misclassified, a significant amount that could have benefited from abstaining from a prediction at all.

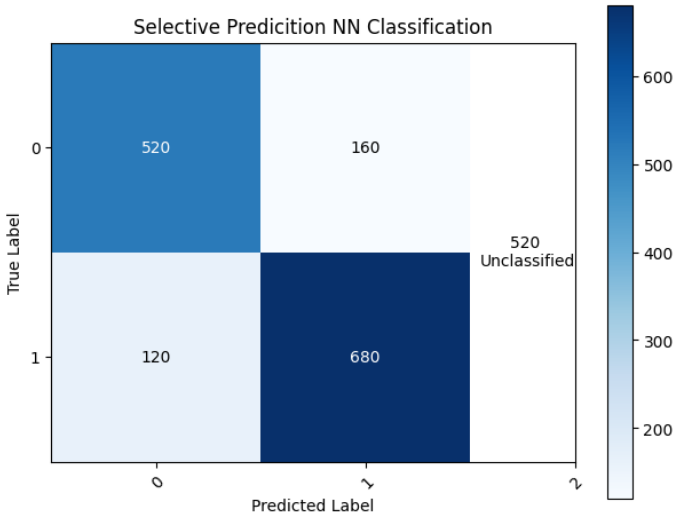


Figure 4: Confusion Matrix for Selective Prediction Neural Network Classification. The neural network with the addition of selective prediction is able to abstain from about 25% of the potential files, a preferable choice as compared to the neural network on it's own. The false positives and false negatives were reduced, however, there are still a significant amount that should have been abstained from.

In an example of a false positive from our model, the unstructured data described "a depression in the skull", which leads us to think that some medical terms that are the same word that may be used to describe different conditions could cause the model confusion.

In addition, some of the false positives and false negatives can be explained by our dataset. Since our dataset only includes patients diagnosed with either only depression or only anxiety, two diagnoses that are similar and often comorbid, patient notes for either diagnoses can have similar information and descriptions that make it so the model would have a more difficult time discerning between the diagnoses. For example, one of our false negative results had patient notes containing descriptions of feeling "worried", with worry often being correlated to anxiety. Thus, the model might perform better on patient data from patients who solely exhibit symptoms that pertain to just one of the diagnoses.

While these preliminary observations from the selective prediction model are promising, the calibration of the model needs further work. The number of unclassified/abstained from files in the selective prediction model were quite proportionate to the number of false positive and negative files in the original classifier, indicating that these false positives and negatives may have been picked up and abstained from, but there still remain a significant number of false positives and negatives in the model. This begs the question as to why the model is continuing to misclassify certain files, but additionally abstain from others that would have been correct in the original neural network. Self-supervised learning techniques would be ideal to resolve this issue, readjusting the thresholds as the model sees fit. Even still, the selective prediction model held a higher accuracy than the original neural network, supporting the use of selective prediction in classification studies.

6 Conclusion

As seen from our results, our model supports the preliminary conclusion that utilizing neural networks for unstructured data abstraction is not only possible but holds a high comparison to the preferred model of Supported Vector Machines (Maarseveen T, 2020). The main area that we aim to improve upon is adding the dimension of "selective prediction" that was described in our inspiration paper via a new threshold marker (Akshay Swaminathan, 2024). In addition, we ran the neural network on our original, larger datasets to look at a global and even more unstructured input for the abstraction capabilities of a neural network over an SVM model.

Through this project, we learned more about not only the context of using NLP techniques for industries outside of computer science, but also specific ways in which implementation of various healthcare research such as selective prediction can be customized to fit a dataset with thousands of unstructured text entries. We learned how to effectively compare our own techniques with many previous techniques such as SVM, and be able to plot them to demonstrate graphically.

However, the number of false positives and negatives was high, as well as the percentage we abstained from through selective prediction. Our primary limitation is that our threshold for uncertainty was very wide, meaning that while this model would be effective for diseases which are not similar (e.g. depression and cancer), further work must be done for diseases with some similar traits (e.g. depression and anxiety) to hold the same accuracy. One extension for future work in order to improve upon this limitation is to self-moderate the threshold for uncertainty instead of simply setting a constant, which our model used.

Overall, we believe that our model makes significant progress on selective prediction of unstructured clinical data and will be able to serve as an effective foundation for future work in differentiating between many diseases in order to better diagnose patients.

Note: In regards to team contribution, all three team members contributed an equal amount of time, effort, and collaboration. Nicole wrote the abstract, introduction, relevant work, and mathematical aspect of the approach. Nathan coded the preprocessing of the dataset as well as section 3.2 and Experiments section of the writeup. Emily trained and tested the model, and worked specifically on the evaluation method and pulled qualitative trends for the Analysis section. All three worked on the model itself and analyzed results as a team.

References

William Wang Ujwal Srivastava Edward Tran Aarohi Bhargava-Shah Janet Y Wu Alexander L Ren Kaitlin Caoili Brandon Bui Layth Alkhani Susan Lee Nathan Mohit Noel Seo Nicholas Macedo Winson Cheng Charles Liu Reena Thomas Jonathan H Chen Olivier Gevaert Akshay Swaminathan, Ivan Lopez. 2024. Selective prediction for extracting unstructured clinical data. In *Journal of the American Medical Informatics Association Volume 31, Issue 1*, pages 188–197, Online. Oxford University Press.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Tushaar Gangavarapu, Aditya Jayasimha, Gokul S. Krishnan, and Sowmya Kamath S. 2020. Predicting icd-9 code groups with fuzzy similarity based supervised multi-label classification of unstructured clinical nursing notes. volume 190, page 105321.

Irene Li, Jessica Pan, Jeremy Goldwasser, Neha Verma, Wai Pan Wong, Muhammed Yavuz Nuzumlalı, Benjamin Rosand, Yixin Li, Matthew Zhang, David Chang, R. Andrew Taylor, Harlan M. Krumholz, and Dragomir Radev. 2022. Neural natural language processing for unstructured data in electronic health records: A review. *Computer Science Review*, 46:100511.

Reinders M Knitza J Huizinga T Kleyer A-Simon D van den Akker E Knevel R Maarseveen T, Meinderink T. 2020. Machine learning electronic health record identification of patients with rheumatoid arthritis: Algorithm pipeline development and validation study. In *JMIR Med Inform*.