Rationale Extraction and Crohn’s Disease Detection from Computed Tomography Enterography Reports

by

Jiayi Dai

A thesis submitted in partial fulfillment of the requirements for the degree of

Master of Science

Department of Computing Science
University of Alberta

Abstract

Building predictive models with higher predictive performance is the common pursuit in text classification tasks. In almost all domains of text classification problems, the current state-of-the-art models (e.g., Bi-LSTM and BERT) are based on deep neural networks that learn language representations with deep and sophisticated neural architectures. However, the lack of interpretability limits the real-world applications of these deep models, especially in life-critical domains. The desire for the interpretability of neural networks that aims to provide predictions along with explanations has been rapidly emerging. Rationale extraction, which is one best practice of explainable artificial intelligence (XAI) in building explainable neural classifiers, learns with only instance-level supervision to identify discriminative features as explanations for predictions; it can be applied in the medical domain to provide explainable disease diagnostic predictions. The scope of the dissertation is on rationale extraction and predictive models, with a focus on detecting Crohn's disease from CT enterography radiology reports. Specifically, the work of the dissertation: 1. explores rationale extraction as a tool for knowledge acquisition from CT enterography reports, 2. introduces IBDBERT, an inflammatory bowel disease (IBD)-specific BERT large language model, which achieves the state-of-the-art classification accuracy in detecting Crohn's disease from CT enterography reports in comparison to CNN, Bi-LSTM and both generic and domain-specific BERT models and 3. constructs the first ensemble architecture of rationale extraction by imitating human interaction.

Preface

Regarding the part of the dissertation related to detecting Crohn's disease from CT enterography reports (i.e., Section 2.3, 2.4, Chapter 3 and Appendix A), the study protocol was approved by the Health Research Ethics Board of University of Alberta Institutional Review Board (Pro00093304).

The work in Section 2.3, 2.4 and Chapter 3 is part of a research collaboration with Mi-Young Kim, Reed Sutton, Ross Mitchel, Randy Goebel and Daniel C. Baumgart, which is targeting a journal publication. Chapter 4 of the dissertation has been published as "Interactive Rationale Extraction for Text Classification" in Trustworthy and Socially Responsible Machine Learning at Neural Information Processing Systems (2022) [1] (non-archival) and Australasian Language Technology Association (2022) [2] (in proceedings) which are separately available with the links https://openreview.net/forum?id=zaJsDuwwdlJ and https://aclanthology.org/2022.alta-1.15.

Acknowledgements

First and foremost, I am deeply grateful to my supervisors Randy Goebel and Mi-Young Kim for offering me the precious research opportunity.

Particularly, I would like to thank Mi-Young Kim for the caring and keeping my study on track and thank Randy Goebel for the inspiration on explainable AI and knowledge which continues to motivate me. I would also like to appreciate Osmar Zaiane serving my committee and providing helpful suggestions.

Finally, I would like to thank my family for the unconditional support and all the lovely friends from the Computing Science department of the University of Alberta for the great joy.

This research was funded by AMII and NSERC.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Background

Text classification, which is the problem of categorizing texts into predefined classes or labels, is one of the most used and successful applications of natural language processing. With the help of the rapidly advancing computing power and the large volume of high-quality labeled datasets, deep neural network-based predictive models have achieved the state-of-the-art performance in text classification tasks in various real-world domains (e.g., sentiment analysis, medical diagnosis, topic labeling). Compared with traditional classification models, either linear (e.g., logistic regression, Naïve Bayes, support vector machine) or non-linear (e.g., k-nearest neighbors algorithm, decision trees), deep neural network (DNN) models generally present higher predictive performance. With more and more sophisticated neural architecture designs and language representations (e.g., Bi-LSTM and BERT [3]), their performance continues to improve over the recent years.

However, the practice of applying DNNs, especially in life-critical domains, is limited because of their poor interpretability and explanability as black-box models. Because a neural model might perform well in a certain dataset by exploiting some biased features, such as genders, physician and hospital names, which can be dataset specific and may not be truly important information for real-world decision making, it is essential for humans to understand the reason for a neural model to make a prediction. In other words, the interpretability of a neural model is desired such that a human decision maker can trust the model's predictions. To improve the trustworthiness of DNNs, the study of explainable artificial intelligence (XAI) has been increasingly catching attention [4].

Generally speaking, the target of the interpretability of neural network-based text classifiers is to provide not only classification predictions but also explanations which are expected to give insights on what features from input texts are discriminative or used for the corresponding predictions. The explanations provided by current XAI methods can be described as two types: (1) feature-level importance values and (2) rationales (i.e., subsets of input tokens). Given an input text, methods that provide feature-level importance values (e.g., LIME [5] and SHAP [6]) generate a real number value for each token of the input text, which indicates the token's contribution for a certain class, either in a positive or a negative way; alternatively, methods that provide rationales, which are called rationale extraction [7–10], select a text fragment as a rationale, which is expected to contain important features, from the original input text and then make a prediction solely based on the rationale. Rationale extraction can be viewed as a machine learning method of building intrinsically interpretable neural classifiers which learns to extract important features with only instance-level supervision. This is the XAI method explored in the dissertation.

When it comes to the automated detection of Crohn's disease (CD) from computed tomography (CT) enterography textual reports, the task is to build a binary text classification model which distinguishes between the reports with CD and the reports without CD. As a type of inflammatory bowel disease (IBD), CD has an increasing prevalence worldwide, and it can affect the whole gastrointestinal tract and most commonly affects the terminal ileum and colon [11]. The data used for the detection of CD is the textual reports of CT enterography which is an imaging technique for detailed small bowel visualisation and is often used as an accurate tool for the early diagnosis and assessment of Crohn's disease [12]. Considering the large volume of patients to screen, an accurate and trustworthy automated system for detecting CD is important since it will provide supportive evidence as assistance for physicians to make efficient diagnosis.

1.2 Objective

In the development of statistical predictive models for text classification problems, from simple linear models relying on linear combinations of input words to complex deep neural models optimized by gradient descent in arbitrarily high-dimensional spaces, there seems to be tension between predictive performance and interpretability. With rationale extraction as an example, a rationale extraction model that extracts whole input texts as rationales provides the least interpretability, but, compared with models which only extract small portions of input texts as rationales, it generally presents more accurate predictions by making use of more information.

Despite the known challenges, higher predictive performance and better interpretability are desired when building predictive models, which motivates this dissertation. More specifically, the work of the dissertation mainly focuses on:

1. Exploring rationale extraction as a tool for knowledge acquisition for the detection of CD from CT enterography textual reports which leads to an algorithm for automatically collecting important features and building rule-based classifiers.

2. Improving the predictive performance of existing classifiers (i.e., BERT models) on detecting Crohn's disease from CT enterography textual reports through IBD-specific language model augmentation which results in IBDBERT.

3. Improving the predictive performance of rationale extraction without compromising interpretability by constructing the first ensemble architecture for rationale extraction which imitates the interactive process of humans for problem solving.

1.3 Outline

Chapter 2 briefly discusses rationale extraction as a machine learning method for building explainable neural classifiers and how it might align with the way humans search for important features. Then, this chapter explores rationale-based knowledge acquisition from

CT enterography reports by extracting discriminative word or phrase-level features from rationales and then detecting Crohn's disease based on the discriminative features, which presents competitive predictive performance with the original rationale extraction method but with better interpretability.

Chapter 3 introduces an inflammatory bowel disease-specific BERT large language model (i.e., **IBDBERT**), which is created by augmenting the original BERT with a corpus of human experts' knowledge on IBD and is then fine-tuned for the task of detecting Crohn's disease from CT enterography reports. IBDBERT confirms the effectiveness of subject matter knowledge on augmenting a generic large language model BERT in terms of improving its predictive performance on a related downstream task.

Chapter 4 proposes a rationale dialogue architecture called **Interactive Rationale Extraction for Text Classification** by imitating human interaction for handling disagreements, which is the first ensemble architecture of rationale extraction. The work of this chapter argues that rationale extraction is compatible with the way humans benefit from exchanging reasons and presents that the architecture with rationale dialogue improves the predictive performance from base rationale extraction models, which is achieved without compromising the interpretability and the faithfulness of the rationale-based explanations.

The theme of the dissertation is on improving the interpretability and the predictive performance of neural network-based text classifiers. All the methods introduced in the dissertation, except **Chapter 4**, are performed on CT enterography reports for the task of detecting Crohn's disease. Considering the very limited amount of available CT enterography reports and the cost of human annotations for "gold" labels, the rationale dialogue algorithm introduced in **Chapter 4** is performed on general text classification datasets (i.e., IMDB movie reviews and 20 Newsgroups) because the algorithm is to handle the disagreement cases for rationale extraction models (i.e., a small portion of all cases) and requires a large volume of labeled testing data for effective evaluation.

Chapter 2

Rationale Extraction

2.1 Introduction

Rationale extraction [7–10] is a machine learning method that identifies text segments as rationales to explain classifiers' predictions. In a rationale extraction setup with a general architecture described in Figure 2.1, two neural networks, a generator and a classifier, work jointly: the generator extracts a subset of text, namely a rationale, from the original input text where the rationale is expected to cover discriminative information, and then the classifier makes a prediction solely based on the extracted information. The rationale is then seen as the evidence or explanation for the prediction. Note that this use of rationales can be coupled with any neural networks to build predictive classifiers.

The original selective rationale extraction model was proposed by Lei *et al.* [7]. Their rationale extraction model is trained with only instance-level supervision (i.e., without token or feature-level supervision) and the model itself learns to identify discriminative features. Their model faithfully explains a neural network-based classifier's predictions by jointly training a generator and a classifier with only instance-level supervision. We summarize their work as follows. The generator g consumes the embedded tokens of the original text,

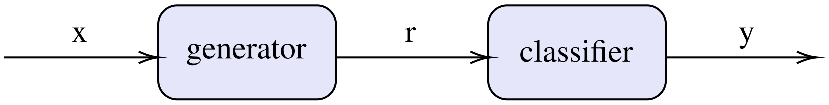


Figure 2.1: The general architecture of rationale extraction. x is an input text, r is a rationale, y is a prediction.

namely $x = [x_1, x_2, ..., x_l]$ where l is the number of the tokens in the text and each token $x_i \in \mathbb{R}^d$ is an d dimensional embedding vector, and outputs a probability distribution $p(z|x)$ over the text mask $z = [z_1, z_2, ..., z_l]$ where each value $z_i \in \{0, 1\}$ denotes whether the corresponding token is selected.

A rationale r is then defined as (z, x) representing the mask z over the original input x. Subsequently, the classifier f takes (z, x) as input to make a prediction $f(z, x)$. Given gold label y, the loss function used to optimize both generator g and classifier f is defined as

$$loss(z, x, y) = ||f(z, x) - y||_2^2 + \lambda_1 ||z|| + \lambda_2 \sum_{i=1}^{l-1} |z_i - z_{i+1}| \tag{2.1}$$

which consists of three parts respectively corresponding to predictive loss, selection loss and contiguity loss. The predictive loss encourages the model to select better discriminative features as rationales and improve predictions. The selection loss and the contiguity loss, respectively fine-tuned by hyper-parameters λ_1 and λ_2, encourage the model to select concise and contiguous rationales, which is intended to improve interpretability. An example of how to compute the selection loss and contiguity loss given a mask can be found below in Figure 2.2.

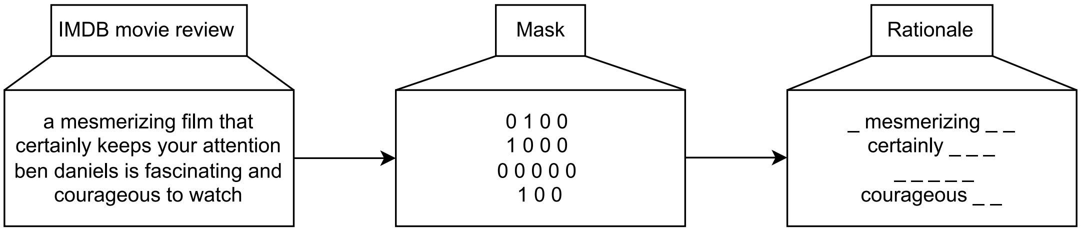


Figure 2.2: Given an IMDB movie review, a mask is created to apply on the original input to produce a rationale. The selection loss is 3 (i.e., the number of 1's in the mask) and the contiguity loss is 6 (i.e., the number of transitions between 0 and 1 along the mask sequence). Note that a mask is composed of discrete 0's and 1's only for inference and a mask is probabilistic during the training process.

Background The original rationale extraction model proposed by Lei *et al.* [7] in 2016 uses hard masking by applying Bernoulli distribution (i.e., non-differentiable) on each token for generating rationales which requires REINFORCE [13] for training. Since REINFORCE does a random sampling of rationales for gradient estimation, this complicates the training process and presents high variance and sensitivity to parameters [9]. Following the select-predict architecture proposed by Lei *et al.* [7], further explored were improved methods for differentiable masking, such as Gumbel-Softmax [14] and HardKuma [8] which respectively approximate the Bernoulli distribution with differentiable Gumbel-Softmax and Hard Kumaraswamy distributions.

Advantages One advantage of rationale extraction is "full faithfulness", which is an important characteristic of model interpretability. An explanation is faithful if the explanation is truly the reason for the prediction [15, 16]. While LIME and SHAP, popular post-hoc methods, and attention weights, for neural networks with attention mechanisms [17–19], can hardly guarantee the features with high importance scores are truly the reason for the prediction, rationale-based explanation is fully faithful since a rationale is the only input for its prediction. Also, once a rationale extraction model is trained, producing each pair of rationale and prediction takes only one inference, which is much more computationally efficient compared to post-hoc processing (e.g., to explain a model's prediction on every single instance, LIME needs to compute a linear model to approximate the behaviour of the model around the instance).

Limitations In the joint training process of rationale extraction proposed by Lei *et al.* [7], optimizing the generator requires the supervisory signal remotely from the predictive loss of the classifier which in turn depends on the rationale selection by the generator. The major challenge of rationale extraction is to train a rationale provider (i.e., a generator) and a classifier jointly with only instance-level supervision where the joint architecture of select-predict is fundamental in order to guarantee the faithfulness of rationales as explanations for the classifier's predictions. Jain *et al.* [9] discussed the problem of the joint training by focusing on the differentiability problem and they showed that a rationale extraction model's predictive performance can be improved by breaking the joint architecture. In their work, the task of a generator selecting tokens is done by applying simple heuristics based on feature important scores (e.g., extracting tokens with the top k importance scores given by any importance scoring method, such as LIME [5], attention scores [20] and input gradients [21, 22] as mentioned in their work) and an independently trained classifier then consumes the selected tokens for prediction. Their work keeps the select-predict architecture (i.e., to remain faithful) and avoids training the two base models jointly.

2.2 Example: Rationale Extraction vs. A Superstitious Hockey Fan

Day	What I drink, eat, wear			Game Result
1	milk	sandwich	yellow	loss
2	juice	sandwich	yellow	loss
3	milk	burger	yellow	loss
4	milk	sandwich	blue	win

Table 2.1: A superstitious hockey fan trying to help his team win by searching in his daily routine for the "cause" of the game result. The features are {what he drinks, eats, wears} and the corresponding options separately are {milk, juice}, {sandwich, burger} and {yellow, blue}.

Rationale extraction is all about searching for discriminative information. Since the search for discriminative information for prediction is common in human activities, we may use the following example to intuitively understand rationale extraction. In the scenario of a superstitious hockey fan (Table 2.1), the hockey fan believes the game result of his favorite hockey team is decided by his actions and he tries to find the variable in his daily life that causes his team to win or lose. Here the hockey fan assumes that the three features (i.e., what he drinks, eats and wears) are the potentially important features and only one feature affects the game result. His strategy is to change one variable at a time and observe if the game result changes. For example, from Day 1 to Day 2, he changes what he drinks from milk to juice while keeping other variables unchanged and he observes that the game result is unchanged. By repeating the process, after four days of searching and observing, the hockey fan believes that his team wins when he wears blue.

	Rationale Extraction	Hockey Fan
Original Data	free texts	tabular
Features for Prediction	subsets of texts	combinations of feature options
Prediction Model	neural classifier	rule: result = win if feature = x
Search Method	gradient descent	variable control
Supervision	instance-level	

Table 2.2: Comparison between rationale extraction and the superstitious hockey fan. In a rule of the hockey fan, we have features $\in$ {what he drinks, eats, wears} and x $\in$ {milk, juice} or {sandwich, burger} or {yellow, blue} corresponding to each feature (as shown in Table 2.1). As an example rule, the hockey fan believes that his team wins if what he wears is blue. The gradient estimation calculations for rationale extraction based on REINFORCE [13] for hard masking and reparameterization trick for differentiable masking can both be found in the work by Jang *et al.* [14].

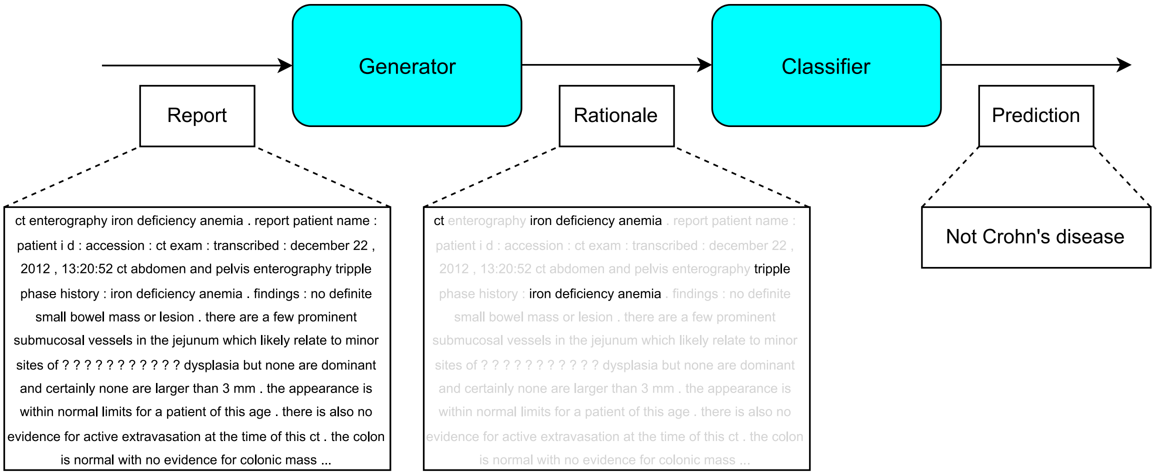
This scenario is comparable to rationale extraction in terms of searching for important features (Table 2.2). While, in rationale extraction, a neural network-based generator selects features from free texts and a neural classifier predicts by consuming the features, the hockey fan does variable control over the three features and makes a prediction based on a simple rule. The hockey fan's belief surely will fail when more games happen, which is determined by the insufficient feature space and the wrong predictive model (i.e., his wrong assumptions). To his defense, the problem of searching for discriminative features in reality is generally difficult to tackle considering the massive amount of features to observe and the potentially complicated interaction within features. However, the superstition and the search attempt could be avoided if he has some knowledge about what features (i.e., probably not what he drinks, eats, wears) truly are potentially important for the problem of predicting game results where knowledge can serve as feature-level supervision.

The superstition of the hockey fan can also exist with predictive models in general, which means what a predictive model considers to be important may not be the true cause of a result in reality. The reason is that, during the supervised training process, a predictive model usually only has access to a dataset with instance-level supervision, which is similar to the features and game results that the hockey fan can observe during the four days. With rationale extraction explicitly declaring which features are used for a prediction, a human decision-maker can figure out if the prediction is probably a result of “superstition”.

2.3 Rationale Extraction for Crohn’s Disease Detection

Canada’s vast geographic area and low population density pose profound challenges for access to highly specialized health care for remote and rural residents. Only 2.4% of all specialists practice in rural and small-town Canada according to the Canadian Institute of Health Information. Rural patients need to travel far, often $>$ 500 km, for access to a specialist and even farther for an IBD expert [23]. An automated diagnostic prediction system for Crohn’s disease, which is trustworthy and explainable, will provide timely diagnostic suggestions and reduce the cost of accessing specialists, especially for patients from remote areas. An automated and interpretable predictive model that detects Crohn’s disease from CT enterography reports might provide some insights for creating such an automated diagnostic system.

The question of interpretability of machine learned predictive classification outputs is a very general challenge that is at the centre of all current research on XAI systems [4]. One best practice in building an explainable classifier is natural language rationale extraction, which can be applied to the detection of Crohn’s disease from CT enterography textual reports to support diagnostic predictions with rationales. The schematic of rationale extraction for the diagnosis of Crohn’s disease is shown in Figure 2.3.



ct enterography iron deficiency anemia . report patient name : patient i d : accession : ct exam : transcribed : december 22 , 2012 , 13:20:52 ct abdomen and pelvis enterography tripple phase history : iron deficiency anemia . findings : no definite small bowel mass or lesion . there are a few prominent submucosal vessels in the jejunum which likely relate to minor sites of ? ? ? ? ? ? ? ? ? ? ? dysplasia but none are dominant and certainly none are larger than 3 mm . the appearance is within normal limits for a patient of this age . there is also no evidence for active extravasation at the time of this ct . the colon is normal with no evidence for colonic mass ...

ct enterography iron deficiency anemia . report patient name : patient i d : accession : ct exam : transcribed : december 22 , 2012 , 13:20:52 ct abdomen and pelvis enterography tripple phase history : iron deficiency anemia . findings : no definite small bowel mass or lesion . there are a few prominent submucosal vessels in the jejunum which likely relate to minor sites of ? ? ? ? ? ? ? ? ? ? ? dysplasia but none are dominant and certainly none are larger than 3 mm . the appearance is within normal limits for a patient of this age . there is also no evidence for active extravasation at the time of this ct . the colon is normal with no evidence for colonic mass ...

Figure 2.3: Schematic of rationale extraction applied to automated detection of Crohn's disease from CT enterography reports. After receiving a CT enterography report, the generator selects rationales (i.e., discriminative text features from the report) as evidence, which is then consumed by the classifier to make a prediction (i.e., either Crohn's disease or not).

2.3.1 Experimental Setups

Base neural networks and training CNN and Bi-LSTM are used as the base neural networks for the generators and the classifiers in rationale extraction. CNN had filter sizes of [3, 4, 5] and 100 filters were used for each filter size. Bi-LSTM had 1 hidden layer with a dimension of 32. For both CNN and Bi-LSTM, the number of training epochs was 30; the dropout rate was set to be 0.2; the batch size was 128; Adam [24] was used as the optimizer with a weight decay rate of 5e-6 and a learning rate of 1e-3; GloVe [25] of 100-dimensional word embedding was used. Note that any black box neural models can serve as a generator and a classifier in a rationale extraction setup.

For all the rationale extraction models, the experiments were repeated five times over five distinct random seeds (i.e., [2022, 2023, 2024, 2025, 2026]) to produce averaged values of predictive performance. In each experiment, when each training epoch of the 30 epochs finishes, the cross-entropy loss of the model on the developing dataset was computed as predictive loss. In the total 30 epochs, the learning rate was halved if there was no improvement in the predictive performance on the developing dataset after every 5 epochs. The version of the model among the training epochs that achieved the best performance on the developing dataset was taken as the final version. The final version was then used for inference on the testing dataset to obtain predictive accuracy. Details about the different splits (i.e., training, developing and testing) of the CT enterography reports can be found in Appendix A.

Rationale masking While a classifier outputs the probability distribution for class prediction, for each input token, a generator was implemented to output the probability distribution for predicting if the token was selected or not (i.e., masking). Gumbel-Softmax [14] was used as differentiable masking, which is for simplifying gradient estimation. For Gumbel-Softmax, the initial temperature was set to be 1 with a decay rate of 1e-5. The implementation of rationale extraction with Gumbel-Softmax was adopted and modified from https://github.com/yala/text_nn.

Hyper-parameters The two hyper-parameters (λ_1, λ_2) from the loss functions 2.1 were set to be [(8e-6, 0), (1e-5, 0)] and [(8e-4), (1e-3, 0)] separately corresponding to rationales lengths of [16, 20] and [17, 20] (as in Table 2.3) for CNN and Bi-LSTM-based rationale extraction models.

2.3.2 Evaluation

In addition to predictive accuracy, we report the average numbers of words in rationales as a proxy for their interpretability (e.g., rationales containing whole reports are least interpretable). It has been reported that, for datasets of various domains (e.g., movie reviews and news), a rationale extraction model's predictive performance increases when its rationale length increases [1, 9], which is also observed for detecting Crohn's disease from CT enterography reports.

In our experiments, the two rationale extraction models that use convolutional neural network (i.e., CNN) [26] and bi-directional long short-term memory (i.e., Bi-LSTM) [27] present similar performance when using generated rationales of similar length. For exam-

Base Model	Predictive Accuracy	Interpretability
CNN	.80, .81	16, 20
Bi-LSTM	.75, .81	17, 20

Table 2.3: Rationale extraction models' predictive accuracy with corresponding rationale lengths for the detection of Crohn's disease from CT enterography reports.

ple, when the hyper-parameter for selection loss is tuned to generate rationales of 20 words on average, both models have a predictive accuracy of 81% as shown in Table 2.3. However, compared with the CNN, Bi-LSTM-based rationale extraction models show a much higher variance in terms of accuracy and rationale length. During the experimental runs of the same hyper-parameter settings (i.e., for selection and contiguity loss), the Bi-LSTM-based model sometimes selects very few words as rationales and shows poor prediction performance that is close to random guessing. For example, over the 5 experiments for the Bi-LSTM-based rationale extraction model with the hyper-parameters $\lambda_1 = 0.001$ and $\lambda_2 = 0$, we obtained the averaged rationale length of [14.75, 25.57, 19.81, 24.73, 0.04] in words with corresponding predictive accuracy of [0.83, 0.84, 0.79, 0.82, 0.48] where rationales of 0.04 words resulted in a predictive accuracy of 0.48 which is similar to random guessing in the binary classification task of Crohn's disease prediction.

Using human knowledge in rationales Attempts that combine human knowledge to directly augment rationales have been conducted, such as augmenting rationales with lost negation information and adding human annotated rationales to machine rationales (i.e., rationales provided by a rationale extraction model) for a trained rationale extraction classifier to make predictions, which did not improve the rationale extraction model's predictive performance.

It has been observed that sometimes entities that are negated in the original texts are no longer negated in the rationales provided by rationale extraction models. Intuitively, negation information is very important for semantics. For example, "no evidence of Crohn's disease" has an opposite meaning towards "evidence of Crohn's disease". However, the text segments considered by humans to be important for semantics may not also be important for a classifier in a rationale extraction setup. When the evidence of Crohn's disease truly exists, a radiologist, instead of writing "there is evidence of Crohn's disease", would use more detailed descriptions for the "evidence" (e.g., mural thickening), which means "evidence of Crohn's disease" almost always appears within a negation environment. In other words, "evidence of Crohn's disease" is as strong as "no evidence of Crohn's disease" in terms of decision making for a rationale extraction classifier and the negation information is not a necessity for the classifier.

The reason that augmenting rationales by direct manipulation (e.g., with negation and human rationale annotations) may not improve the predictive performance of a rationale extraction model ultimately lies within the training data and the training process. Because the classifier in rationale extraction is trained with the rationales extracted from its training data, what the classifier considers to be important is constrained by the training data. More discussions about the relationship between what a model considers to be important and the training process can be found in Subsection 2.2. However, it has been reported that providing human rationale annotations as token-level supervision during the training process is helpful in terms of improving the interpretability of machine-generated rationales [28] and might slightly improve the rationale extraction model's predictive performance in some tasks [29].

2.4 Rationale-based Knowledge Acquisition from CT Enterography Reports

Compared to post-hoc processing, which searches for what a learned model considers to be important, rationale extraction searches for important features directly from a dataset, which is more suitable to support knowledge acquisition. The task of knowledge acquisition here is to automatically collect features that are important for detecting Crohn's disease from CT Enterography textual reports (i.e., strong discriminative indicators) and then make uses of the important features to create predictions using rules formulated from those indicators.

2.4.1 Strong Indicators from Rationales

From the rationales generated by the rationale extraction model, we can furthermore discover strong phrase or word-level indicators which provide straightforward insights for diagnostic predictions. The process of automatically discovering strong indicators is described in Figure 2.4.

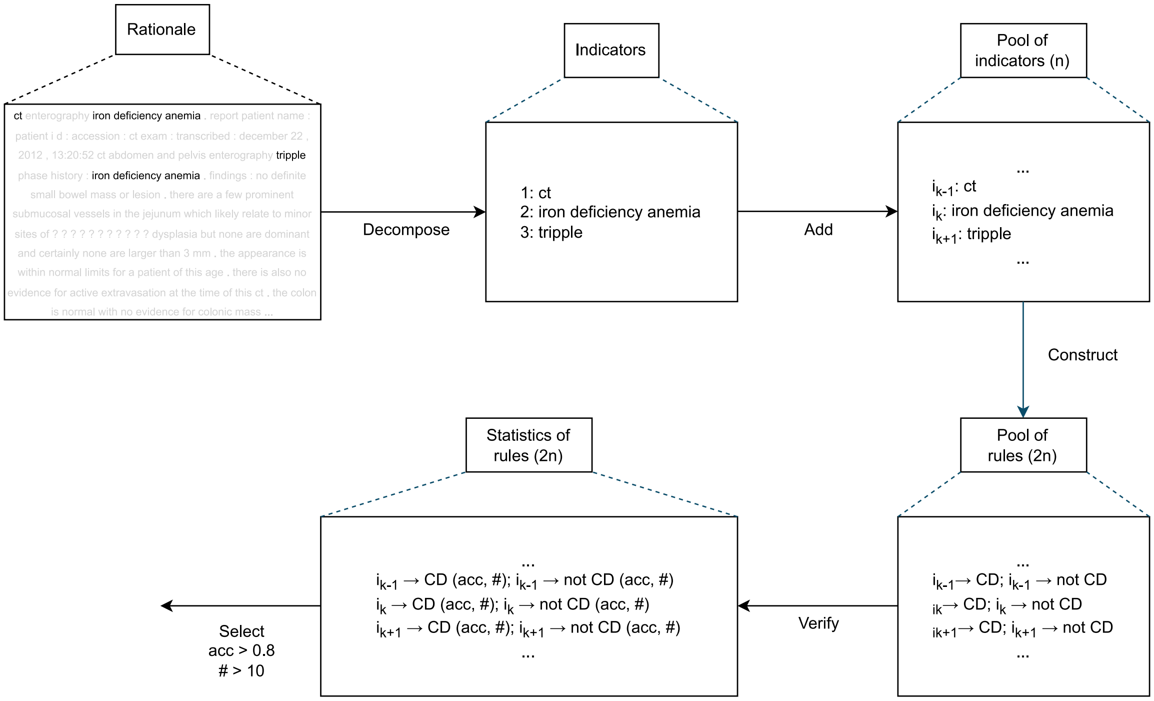

Rationale
ct enterography iron deficiency anemia . report patient name : patient i d : accession : ct exam : transcribed : december 22 , 2012 , 13:20:52 ct abdomen and pelvis enterography tripple phase history : iron deficiency anemia . findings : no definite small bowel mass or lesion . there are a few prominent submucosal vessels in the jejunum which likely relate to minor sites of ? ? ? ? ? ? ? ? ? ? ? dysplasia but none are dominant and certainly none are larger than 3 mm . the appearance is within normal limits for a patient of this age . there is also no evidence for active extravasation at the time of this ct . the colon is normal with no evidence for colonic mass ...
Decompose
Indicators
1: ct
2: iron deficiency anemia
3: tripple
Add
Pool of indicators (n)
...
i_{k-1}: ct
i_k: iron deficiency anemia
i_{k+1}: tripple
...
Construct
Pool of rules (2n)
...
i_{k-1}→ CD; i_{k-1} → not CD
ik→ CD; i_k → not CD
ik+1→ CD; i_{k+1} → not CD
...
Verify
Statistics of rules (2n)
...
i_{k-1} → CD (acc, #); i_{k-1} → not CD (acc, #)
i_k → CD (acc, #); i_k → not CD (acc, #)
i_{k+1} → CD (acc, #); i_{k+1} → not CD (acc, #)
...
Select
acc > 0.8
# > 10

Figure 2.4: Schematic of identifying strong predictive indicators from decomposing rationales. For indicator i_k, "$i_k \rightarrow$ CD" refers to the rule "if i_k exists in a report, then the report is of CD"; (acc, #) = (accuracy, number of occurrences) of the rule in the training dataset. As an example, the rule "if *iron deficiency anemia* exists, the report is *Not CD*" has a predictive accuracy of 90% in the 73 out of 1,568 training reports containing the phrase *iron deficiency anemia*.

Given a rationale, we first decompose it into disjoint phrases or words which can be viewed as potentially strong indicators. By applying the decomposition step for all the reports used for training, we obtain a pool of all indicators (say n of them). For each indicator i_k in the pool, two simple rules can be created: if i_k exists in a report, the report is Crohn's disease; if i_k exists in a report, the report is not Crohn's disease, which gives us a pool of rules ($2n$ of them). The reason for constructing the second rule, which seems to be redundant and contradicting the first rule, is that rationale extraction has been observed to sometimes extract discriminative features for wrong classes (i.e., failing to incorporate context). Each rule makes a diagnostic prediction based on the existence of some indicator. After verifying the rules on the labeled training reports, we can measure the predictive performances of all the rules together with their numbers of occurrences in the training reports. We then select top indicators by applying a filter on the rules based on their performances and occurrences (i.e., predictive accuracy $> 80\%$ and occurrences > 10 in our experiments).

2.4.2 Rule-based Classifier Using Strong Indicators

In a rationale extraction model, a classifier makes predictions by consuming the rationales extracted by its generator. So, if the generator fails to provide a good rationale that contains important information for some specific report, the classifier may not be capable of making a correct prediction. However, the top indicators might help in cases where a rationale extraction model can fail because the top indicators may be verified to be working well globally in the overall training reports.

With that motivation, we have also constructed a companion rule-based classifier by using the automatically discovered strong indicators from the previous subsection. The rule-based algorithm is designed to make predictions by simply comparing the numbers of the occurrences of top indicators for CD and not CD (e.g., if a report contains more top indicators for CD than not CD, the report is predicted as CD). Formally, given the collections of top indicators separately for CD and not CD (say I^+, I^-) and an input report

x, we have

$$f(x, I^+, I^-) = \sum_{i \in I^+} \delta(i, x) - \sum_{i \in I^-} \delta(i, x)$$

where $\delta(i, x)$ is a binary value denoting the existence of indicator i in report x (i.e., $\delta(i, x) = 1$ if i is in x; $\delta(i, x) = 0$ if i is not in x). The final prediction is defined by the conditional function

$$pred(x, I^+, I^-) = \begin{cases} CD, & f(x, I^+, I^-) > 0 \\ not\, CD, & f(x, I^+, I^-) < 0 \end{cases}$$

In the cases where this algorithm does not apply (i.e., $f(x, I^+, I^-) = 0$), we then use the original rationale extraction model's predictions.

Experimental setup The CNN-based rationale extraction model was preferred for identifying strong indicators since we observed that it provides more sparse rationales compared with Bi-LSTM. This helps inform rationale decomposition, which aids the observer in understanding rationale components and their semantic relationship. With the CNN-based model, which achieves 81% accuracy with 20-word rationales in the overall testing reports as the rationale provider, we obtain a rationale rule-based classifier. The experiments for the rationale-rule model were repeated five times along the five experiments for the CNN-based rationale extraction model (i.e., the CNN-based rationale extraction model in each experiment was used to provide rationales to be decomposed by the rationale-rule model).

The rationale extraction model details can be found in Section 2.3.1. The process of discovering strong indicators was applied in the training data and then the testing data was used for evaluation. The details about the dataset splits can be found in Appendix A.

Performance In the averaged 158.8 cases of the total 198 testing reports where the rules apply, the purely rule-based algorithm achieves a predictive accuracy of 84.1%, while the rationale extraction model's accuracy is 84.0%. In the complete set of 198 cases, the rule-based classifier achieves an accuracy of 81.4%, compared with the accuracy of 81.3% for the original rationale extraction model.

Compared with the rationale extraction model based on a black-box neural network-based classifier, a rule-based classifier provides clearer explanations and allows human experts to investigate and modify the rules. In the 158.8 cases where the algorithm applies, the predictions only depend on strong indicators of 4.5 words which is much more straightforward than the rationales of 18.3 words. Also, the rules of the strong indicators allows human experts to refine and enhance. For example, the term *humira* has been discovered as a strong indicator for Crohn's disease (i.e., in the 35 training reports containing the term, the rule "*humira* $\rightarrow$ CD" achieves 100% predictive accuracy) and human experts might enhance the rule by considering the real-world cases where the rule can possibly fail. In addition, as shown in the experimental results, the rule-based classifier using the strong indicators extracted from the rationale extraction model is competitive with the original rationale extraction model in terms of predictive performance.

Limitations When a set of tokens is selected as a rationale by a neural generator, the rationale is supposed to be treated as a whole which should be viewed as a contextualized summary of the original text and can also involve the interaction among the subset tokens. However, decomposing a rationale into independent tokens and phrases might cause a loss of context and details of interaction. Similarly, the algorithm, which takes use of the strong indicators through simply counting the numbers of appearances of the strong indicators for both labels, is linear to the independent existence of the strong indicators and might also ignore the potentially useful interaction among the indicators.

Chapter 3

IBDBERT: An Inflammatory Bowel Disease-specific BERT Model

3.1 Introduction

Our initial text analysis of the radiology reports on existing data have used several artificial intelligence and machine learning (AI/ML) tools, including images-based neural models, i.e., convolutional neural networks (CNN) [26] and bi-directional long short-term memory neural networks (Bi-LSTM), i.e. a bi-directional version of LSTM [27]. In addition to those established neural network predictive modeling methods of CNN, Bi-LSTM, we have also considered the large language model (LLM) BERT [3], a distilled version of BERT, DistilBERT [30], and a domain specific variant, BioClinicalBERT [31]. All of these are evaluated for their classification performance in comparison to our own creation of a domain-specific LLM which we call IBDBERT.

The Bidirectional Encoder Representations from Transformers model (BERT) [3] is one of the first large language models that uses the text encoder from the original Transformer model [32]. A BERT model can be pre-trained on an unlabeled corpus (e.g., the original BERT was trained on BookCorpus [33] and English Wikipedia) using a masked language modeling for the language representation, and then fine-tuned on labeled data for a downstream domain-specific classification task. BERT has been widely adopted in medicine for domain-specific pre-training (e.g., MedBERT [34], BioClinicalBERT [31] and BEHRT

[35]) and related text classification tasks [36, 37], and has achieved the state-of-the-art performance.

Masked language processing is implemented by providing the model with texts containing a portion of randomly masked words and then, based on how well the model predicts the masked words, adjusting the appropriate weights within the model. The task of predicting a masked word is implemented as a classification process over the whole BERT vocabulary (i.e., around 30,000 words for the original BERT) and cross-entropy is used as prediction loss to refine the BERT network. Overall, this LLM construction method provides the base LLM called BERT.

3.2 Developing IBDBERT for Crohn's Disease Detection

Inflammatory bowel disease-specific BERT (IBDBERT) is developed by augmenting the original BERT with an inflammatory bowel disease-related textbook [38] and guidelines [39–60] through a masked language modeling, which does not require any extra annotations or labelling (i.e., unsupervised pre-training). IBDBERT is then fine-tuned through supervised learning with labeled CT enterography textual reports for the binary classification task of detecting Crohn's disease. The overall process of developing IBDBERT and applying IBDBERT in Crohn's disease detection is described in Figure 3.1.

The purpose of the augmentation is to extend the original BERT model with IBD-related language without diminishing the general language understanding captured by the original BERT. Our development of IBDBERT is similar to the work of BioClinicalBERT, BEHRT and MedBERT, where the training process all follows masked language modeling of the original BERT. However, BioClinicalBERT, BEHRT and MedBERT were all trained on data directly collected from patients (i.e., separately MIMIC-III [61], Clinical Practice Research Datalink [62] and general electronic health records) while our IBDBERT was trained on human experts' knowledge on inflammatory bowel disease. Our corpus used for training IBDBERT contains around 0.9 million words, which is a relatively small dataset compared

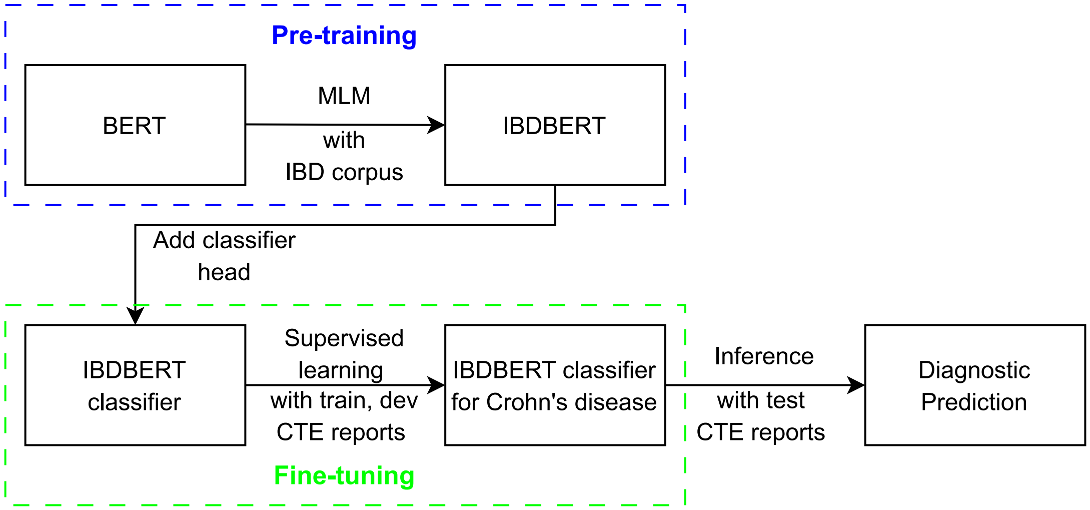


Figure 3.1: Developing IBDBERT for Crohn's disease diagnostic prediction: pre-training BERT through masked language modeling with IBD corpus and fine-tuning through supervised learning with Crohn's disease classification on CT enterography reports.

to 3,300 million words used for the original BERT and the extra data collected from 1.6 million and 20 million patients for BEHRT and MedBERT. But our motivation is to confirm whether subject matter expert knowledge can be coupled with generic LLMs like BERT, to improve predictive classification performance.

3.3 Experimental setup

3.3.1 Masked Language Modeling for IBDBERT

Since the BERT pre-training process by masked language modeling (MLM) is not fully deterministic in distributed computing, the pre-training process given each learning rate (LR) was repeated three times to reduce variance. The LR options experimented with were [5e-5, 4e-5, 3e-5, 2e-5, 1e-5, 5e-6, 1e-6].

In implementation, the block size was 256 and the batch size was 8 which means every 256 tokens were in one block as input for MLM and one batch consisted of 8 blocks. For masked language modeling, where selected words were removed from the blocks, the probability of each word being masked was 15% which means overall 15% tokens were masked in every block of 256 tokens. The maximum number of steps for each pre-training run was 5,000. All other parameters were set to their defaults. The code for MLM was adopted from the publicly available repository from Hugging Face[1].

3.3.2 Fine-tuning for Crohn's Disease Detection

After each pre-training run of one LR option for MLM, the fine-tuning process for classification was repeated five times for an average predictive performance (i.e., a total of 3 pre-training * 5 fine-tuning = 15 classifications were performed for each pre-training LR).

The architectures of the classifiers using BERT models, including the original BERT (bert-base-uncased) [3], DistilBERT [30] and BioClinicalBERT [31], were all their published versions[2] and the default configurations were used. When training all BERT models as classifiers (i.e., the fine-tuning process), the batch size was 8 and Adam optimizer [24] with default parameters was used except the learning rate being specified to be 1e-5.

3.3.3 Training Neural Classifiers

For all the neural classifiers used for the task of detecting Crohn's disease from CT enterography reports, including CNN, Bi-LSTM and all the BERT models including IBDBERT, the experiments were repeated five times over five distinct random seeds (i.e., [2022, 2023, 2024, 2025, 2026]) to produce averaged values of predictive performance. The CNN and Bi-LSTM-based classifiers followed the same training strategy (i.e., the number of training epochs and the selection of the final model) as when they were used in rationale extraction models (see Section 2.3.1). The BERT family models (i.e., BERT, DistilBERT, BioClinicalBERT and IBDBERT) all used 4 training epochs and the version that achieved the lowest cross-entropy loss in the developing data was selected for inference on the testing data. The details about the CT enterography dataset can be found in Appendix A.

[1] https://github.com/huggingface/notebooks/blob/main/examples/language_modeling.ipynb

[2] The BERT models are available by searching the model names on https://huggingface.co.

3.4 Evaluation

During the development of IBDBERT, we observed that the learning rate (LR) for pre-training has a significant effect on the predictive performance of its downstream classification task of detecting CD from CT enterography reports. Intuitively, the LR controls the rate at which training weights are changed based on classification errors, so a higher LR means faster change across epochs of training. Among all the LR values we have experimented with (i.e., [5e-5, 4e-5, 3e-5, 2e-5, 1e-5, 5e-6, 1e-6]), IBDBERT achieves the highest classification accuracy (i.e., 88.5%) when the pre-training LR is 1e-5. When the LR is smaller or larger than 1e-5, the performance overall separately increases or decreases as the LR increases (i.e., the peak performance is achieved by LR = 1e-5). The result is reasonable as a larger LR might cause BERT to "forget" its previous general language understanding and a small LR might cause BERT to be unable to learn much from the added IBD knowledge. The detailed performance of IBDBERT over the different values of LR is reported in Figure 3.2.

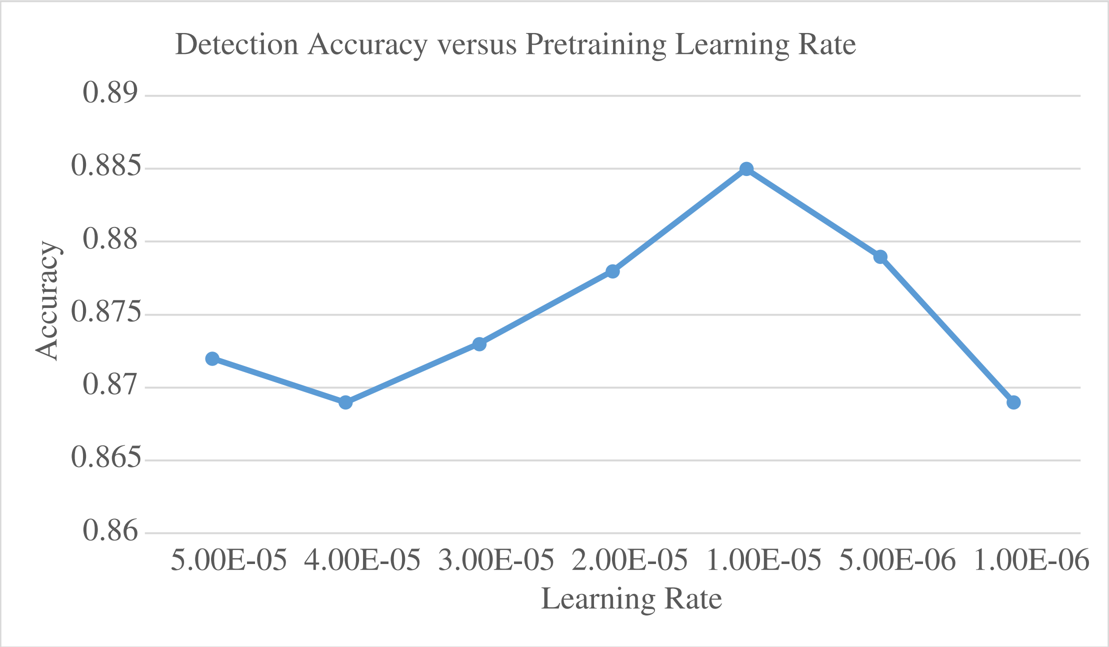


Figure 3.2: IBDBERT performance. We report how the predictive accuracy of IBDBERT changes over different options of learning rate for pre-training.

In terms of predictive performance and interpretability, we compare IBDBERT with rationale extraction models (i.e., based on CNN and Bi-LSTM) and other classifiers which use rationale-rules (from Section 2.4.2), CNN, Bi-LSTM, the original BERT, DistilBERT and BioClinicalBERT. Generally, under multiple LR settings, IBDBERT achieves a predictive accuracy of around 88% in detecting CD; this outperforms the original BERT by 2% and BioClinicalBERT by 3%. The summarized experimental results for all models' performance are reported in Table 3.1. While the rationale-rules classifier is most interpretable due to the self-explainable nature of rules, IBDBERT achieves the highest predictive accuracy.

Task	Model	Predictive Accuracy	Interpretability
Rationale Extraction	CNN	.80, .81	16, 20
	Bi-LSTM	.75, .81	17, 20
Classification	Rationale-Rules	.80, .81	Mostly self-explainable
	CNN	.84	N/A
	Bi-LSTM	.85	N/A
	BERT	.86	N/A
	DistilBERT	.86	N/A
	BioClinicalBERT	.85	N/A
	IBDBERT	.88	N/A

Table 3.1: Predictive accuracy of all methods on detecting CD from CT enterography reports. Accuracy = correct predictions/total number of testing cases. For rationale extraction methods, we report their performances corresponding to their rationale lengths. For example, the CNN-based rationale extraction model achieves an accuracy of 81% when selecting rationales of 20 words. The average lengths of rationales (i.e., numbers of words) are reported as a proxy to rationale interpretability. The interpretability of the neural classifiers is marked as "N/A" meaning that the models do not provide an explanation about which part of a report is discriminative for a prediction.

Chapter 4

Interactive Rationale Extraction for Text Classification

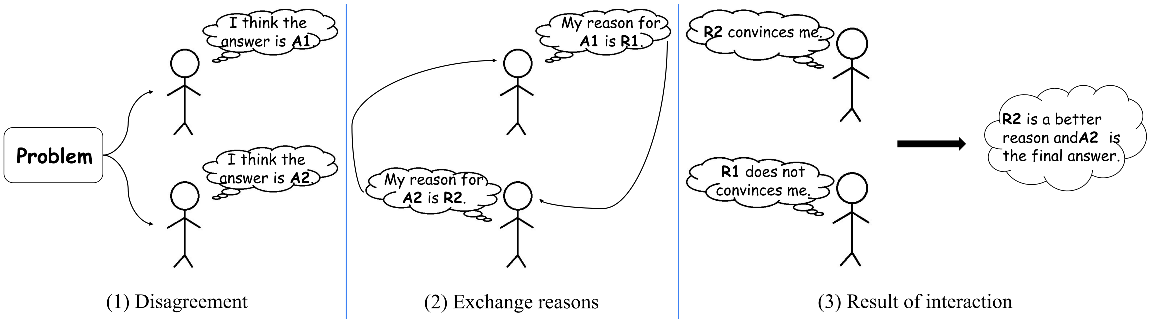
4.1 Introduction

Deep neural networks show superior performance in text classification tasks, but their poor interpretability and explainability can cause trust issues. For text classification problems, the identification of textual sub-phrases or "rationales" is one strategy for attempting to find the most influential portions of text, which can be conveyed as critical in making classification decisions. Selective models for rationale extraction faithfully explain a neural classifier's predictions by training a rationale generator and a text classifier jointly: the generator identifies rationales and the classifier predicts a category solely based on the rationales. The selected rationales are then viewed as the explanations for the classifier's predictions. Through exchange of such explanations, humans interact to find more trusted explanations and achieve higher performance in problem solving. To imitate the interactive process of humans, we propose a simple interactive rationale extraction architecture that selects a pair of rationales and then makes predictions from two independently trained selective models. We show how this architecture outperforms both base models for text classification tasks on datasets *IMDB movie reviews* and *20 Newsgroups* in terms of predictive performance.

Selective (or select-predict) models for rationale extraction in text classification [7, 8],

[0] The implementation is provided on https://github.com/JiayiDai/RationaleExtraction.

with the general structure shown in Figure 4.2a, are designed to extract a set of words, namely a *rationale* [63], from an original text. For prediction purposes, the rationale is expected to be *sufficient* as the input for the classification model to obtain the same prediction based on the whole text. For the purpose of interpretability, the rationale should be *concise* and *contiguous*. A rationale extraction model is *faithful* if the extracted rationales are truly the information used for classification [15, 16]. The problem of extracting rationales that satisfy the criteria above is complex from a machine learning perspective and becomes more difficult with only instance-level supervision (i.e., without token-level annotations) [9]. One model's identification of rationales can suffer from high variance because of the complex training process. An ensemble of more than one model helps to reduce variance, which leads to the exploration of *how to make use of two rationale extraction models and how to make a choice when the two models make different predictions*.


Problem
I think the answer is A1.
I think the answer is A2.
My reason for A1 is R1.
My reason for A2 is R2.
R2 convinces me.
R1 does not convinces me.
R2 is a better reason andA2 is the final answer.

(1) Disagreement (2) Exchange reasons (3) Result of interaction

Figure 4.1: A scenario of the interaction between two students solving a mathematical proof: they disagree with each other, exchange reasons and reach a common conclusion.

When two humans have different answers to a problem (see Figure 4.1), they tend to exchange their reasons or explanations, after which there might be a change of mind. To show why this interaction of humans is effective, we use the problem of proving a mathematical conjecture as an instance: because searching for a correct mathematical proof, which then leads to a correct claim about the conjecture, is usually much more difficult than verifying a proof (e.g., $\mathcal{P} \subseteq \mathcal{NP}$ in computation theory), often one who is not capable of finding a good proof can tell if a proof is good when the proof is given. Considering the complexity for a generator to search among all possible rationales with only remote instance-level supervision, the work of rationale extraction can be much more difficult than classification.

We may then consider selective models for rationale extraction to be naturally compatible with the interactive pattern of humans by viewing the rationales extracted by a generator as the proof for the decisions of its classifier, which means the interaction between two base models can be performed by the exchange of their rationales. Subsequently, the problem becomes how to decide if a rationale is good or not so that we know which pairs of rationale and prediction are appropriate choices when two base models make different predictions. A *good rationale* here is expected to give a correct prediction when input to a decent classifier.

Intuitively, a good rationale is supposed to contain strong indicators for the correct "gold label" instead of insignificant words which do not contribute to classification, which leads to two simple rules for handling base models' disagreements:

1. A good rationale is more likely to produce consistent predictions among classifiers (i.e., a good explanation convinces people);

2. A good rationale is more likely to produce a higher *confidence level* (Section 4.3) for the prediction of one classifier (i.e., one with a good reason is often confident).

These two rules create a basis for classification, as opposed to random guessing based on otherwise randomly selected words. Note that the two rules are based on the assumption that the probability that base models extract strong indicators for wrong labels is very low, which should be considered to be true for decent generators and decent classifiers (i.e., better than random guessing).

To imitate the interactive pattern of humans in problem solving, we introduce **Interactive Rationale Extraction for Text Classification** to interactively connect two independently trained selective rationale extraction models. We show that the architecture achieves higher predictive performance than either base models with similar performance on *IMDB movie reviews* and *20 Newsgroups*. This is done by selecting pairs of rationale and prediction from the base models using the above simple rules. In addition, because this interactive

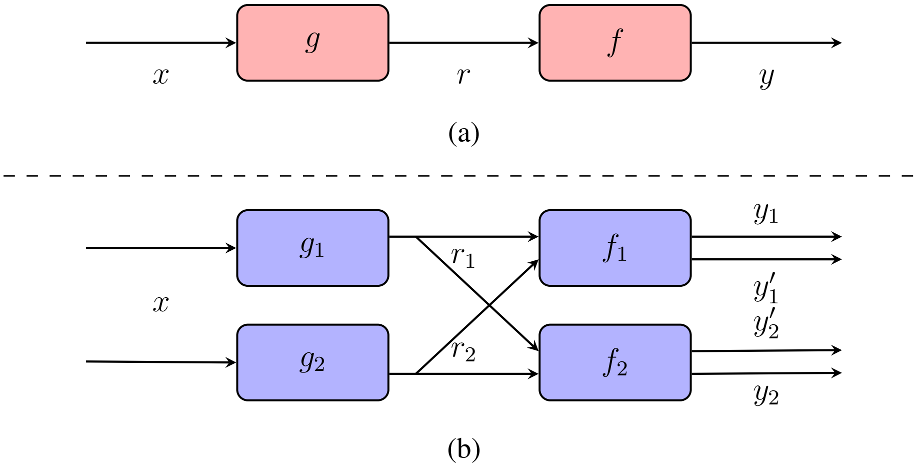


Figure 4.2: (a) Schematic of selective rationale extraction models where x is an embedded text, g is a generator and f is a classifier. Generator g extracts a rationale r based on which classifier f makes a prediction y. (b) Schematic of our interactive rationale extraction where rationales are exchanged.

architecture makes decisions solely based on the base models' rationales, the faithfulness and interpretability of the base models' rationales are not compromised.

4.2 Selective Rationale Extraction

The original selective rationale extraction model was proposed by Lei *et al.* [7] with an architecture shown in Figure 4.2a. Their model faithfully explains a neural network-based classifier's predictions by jointly training a generator and a classifier with only instance-level supervision. The select-predict architecture limits the features consumed by the classifier to be only the rationales extracted by the generator, which guarantees the faithfulness of the rationale-based explanations. The more detailed training process and background information can be found in Section 2.1.

4.3 Confidence Level

Confidence level (CL) indicates how far a neural network's prediction is from being neutral. Given a neural network's non-probabilistic output $k = [k_1, k_2, ..., k_n]$ for a n-class classification, Kumar *et al.* [64] defined the CL of the classification with a Softmax function

$$CL(k) = \frac{exp(max(k))}{\sum_{i=1}^{n} exp(k_i)} \tag{4.1}$$

where $max(k)$ is the highest value among the output nodes $k = [k_1, k_2, ..., k_n]$ and $exp()$ refers to the exponential function. The Softmax function normalizes the non-probabilistic output into a probability distribution over prediction classes, which is used to indicate how far the final prediction is from neutrality.

Guo *et al.* [65] stated that a classification network should not only have a high accuracy but also indicate how likely each prediction is correct or incorrect for trust purposes. In addition, their study on neural networks' calibration [65] suggested that accuracy, even if not nearly identical to CL for some neural networks, is generally positively correlated to CL. This means that, when two base models with similar expected performance make different predictions, the prediction with a higher CL is generally more likely to be correct. Also, a good rationale that contains strong predictive indicators should be more likely to cause a more confident prediction when compared with rationales with less discriminative words (e.g., random words).

4.4 Rationale Dialogue Algorithm

As demonstrated in Figure 4.2b, after the interaction between two base select-predict models, a total of 4 predictions are generated: $y_1 = f_1(r_1)$, $y_1' = f_1(r_2)$, $y_2' = f_2(r_1)$ and $y_2 = f_2(r_2)$ where y_1 and y_2 are the predictions based on their own rationales and y_1' and y_2' are predictions based on the exchanged rationales, as shown in the table below.

	r_1	r_2
f_1	y_1	y_1'
f_2	y_2'	y_2

Given an input text, when the predictions of two base models are the same, namely $y_1 = y_2$, both rationales r_1, r_2 are considered as good and the final prediction is the shared prediction. When two base models initially show a disagreement, we check if one rationale causes more consistent predictions. For example, in the case of two rationale extraction models, consistency means a rationale causes the two models to produce the same prediction. If r_1 causes more consistent predictions, in other words, if r_1 changes the prediction of f_2 to y_1 when given as an input rationale (namely, $y_1 = y_2'$), but r_2 does not change the prediction of f_1 to y_2 when given as an input rationale ($y_2 \neq y_1'$), then the pair (r_1, y_1) is chosen as the final rationale and prediction; symmetrically, if r_2 causes more consistent predictions, the pair (r_2, y_2) is chosen. For the cases where no rationale causes more consistent predictions, we rely on confidence levels which are real numbers between 0 and 1 as defined by Equation 4.1. If the confidence level of f_1 on r_1 is higher than that of f_2 on r_2 (say $CL(f_1, r_1) > CL(f_2, r_2)$ with (f_1, r_1) and (f_2, r_2) separately denoting their corresponding non-probabilistic outputs), the pair (r_1, y_1) is chosen; otherwise, the pair (r_2, y_2) is chosen. The process of selecting a rationale-prediction pair is specified in Algorithm 1. It is worth mentioning that, in implementation, the exchange of rationales only needs to be performed when base models have a disagreement in prediction (i.e., $y_1 \neq y_2$). Also, even though only two base rationale extraction models are used in the algorithm, the basis of the two rules for handling the disagreement cases can be generally applied for an arbitrary amount of base models.

4.5 Rationale Experiments

4.5.1 Datasets

IMDB movie reviews [66] This is a dataset of 50,000 movie reviews collected from the Internet Movie Database (IMDB) with binary labels (i.e., positive and negative). The dataset is originally split into two subsets: 25,000 for training and 25,000 for testing. We randomly split the training data into 20,000 (80%) for training and 5,000 (20%) for development. The numbers of the two labels are perfectly balanced in each subset.

Algorithm 1 Rationale-prediction Selection after Interaction

```
Require: $f_1, f_2, r_1, r_2, y_1, y'_1, y'_2, y_2$ from Figure 4.2b, $CL(f, r)$ for the confidence level of $f$ on $r$.
  if $y_1 = y_2$ then                                         ▷ agreement
      return $(r_1, y_1)$                                     ▷ or $(r_2, y_2)$
  else                                                        ▷ disagreement
      if $y_1 = y'_2$ and $y_2 \neq y'_1$ then                ▷ model 2 convinced by model 1
          return $(r_1, y_1)$
      else if $y_1 \neq y'_2$ and $y_2 = y'_1$ then           ▷ model 1 convinced by model 2
          return $(r_2, y_2)$
      else
          if $CL(f_1, r_1) > CL(f_2, r_2)$ then               ▷ model 1 is more confident
              return $(r_1, y_1)$
          else                                                ▷ model 2 is more confident
              return $(r_2, y_2)$
          end if
      end if
  end if
```

20 Newsgroups It is a publicly available dataset containing a total of 18,846 news articles, with 11,314 for training and 7,532 for testing, in 20 distinct categories of news topics. We split the training data randomly into 9,051 (80%) for training and 2,263 (20%) for development. The numbers of the 20 labels are not perfectly balanced and vary from 304 to 490 in the training data, from 73 to 131 in the development data and from 251 to 399 in the testing data.

4.5.2 Setup

Training Instead of REINFORCE [13], a reparameterization heuristic Gumbel-Softmax [14] is used to simplify gradient estimation. A convolutional neural network [26] is used for both generators and classifiers with filter sizes of 3, 4 and 5, filter number of 100 and dropout rate of 0.5 all following the parameter settings of the original paper. Hidden dimensions of 100 and 120 are separately used for the first and the second base model, which is the only difference among all parameters for training two base models. Adam is used as the optimizer with a weight decay of 5e-06 and an initial learning rate of 0.001. If

20 Newsgroups				
(λ_1, λ_2)	(5e-3, 0)		(1e-3, 1e-3)	
Base Model	Model 1	Model 2	Model 1	Model 2
Length	11.33	11.18	21.76	22.68
Contiguity Loss	17.12	16.84	21.92	21.45
Interaction Cases	(331, 363, 1129, 1211.5) (4.4%, 4.8%, 15.0%, 16.1%)		(228.6, 264, 974.2, 1075.8) (3.0%, 3.5%, 12.9%, 14.3%)	
Case Accuracy	(0.41, 0.43, 0.30, 0.26)		(0.38, 0.44, 0.31, 0.27)	
IMDB movie reviews				
(λ_1, λ_2)	(1e-3, 0)		(2e-4, 2e-4)	
Base Model	Model 1	Model 2	Model 1	Model 2
Length	13.99	17.59	29.22	27.37
Contiguity Loss	21.84	26.45	37.14	35.48
Interaction Cases	(855.6, 946.0, 1187.4, 1250.0) (3.4%, 3.8%, 4.7%, 5.0%)		(681.7, 665.2, 1101.8, 1295.7) (2.7%, 2.7%, 4.4%, 5.2%)	
Case Accuracy	(0.66, 0.65, 0.59, 0.59)		(0.66, 0.64, 0.58, 0.60)	

Table 4.1: Experiment details (average values). We report the rationale length (i.e., number of words) and contiguity loss of each base model and also numbers of interaction cases and each case's accuracy under each hyper-parameter setting. Four values in an interaction case are the average numbers/percentages of the cases separately for base model 1 convinced, base model 2 convinced, base model 1 more confident, and base model 2 more confident. These are the four cases from handling disagreements in Algorithm 1.

no improvement is achieved in loss in development dataset from the previous best model after 5 epochs, the learning rate is halved (i.e., 0.001, 0.0005...) and the training process starts over from the previous best model. In total, 20 epochs are used for training. Cross-entropy is used as the loss objective. Following the setting in the original Gumbel-Softmax paper [14], the initial temperature is 1 with a decay rate of 1e-5. Batch size is set to be 128. GloVe [25] of embedding dimension 300 is used for word embedding[1]. The maximum

[1] Rationale extraction experiments with batch size = 64 and embedding dimension = 100 were conducted and did not show notable difference in base models' predictive performance.

text lengths are separately set to be 80 and 200 words for *20 Newsgroups* and *IMDB movie reviews*.

Testing For each dataset, two base models are independently trained and tested with two settings of hyper-parameters (λ_1, λ_2) for the loss function 2.1. $\{(0.005, 0), (0.001, 0.001)\}$ are used for *20 Newsgroups* and $\{(0.001, 0), (0.0002, 0.0002)\}$ are used for *IMDB movie reviews*. The four settings are chosen in a way to show the performance of the algorithm under different rationale length and contiguity (Table 4.1). For each hyper-parameter setting, both base models are trained and tested with 6 random seeds (i.e., {2022, 2023, 2024, 2025, 2026, 2027}), and the invalid cases where two base models show a significant difference in the performance in development dataset (i.e., $> 3\%$ in accuracy) are removed. The numbers of invalid cases are separately 2, 1, 1, 0 out of 6 for the four hyper-parameter settings.

	20 Newsgroups		IMDB movie reviews	
(λ_1, λ_2)	(5e-3, 0)	(1e-3, 1e-3)	(1e-3, 0)	(2e-4, 2e-4)
Model 1	.55 (.53-.57)	.58 (.56-.59)	.81 (.80-.82)	.82 (.81-.83)
Model 2	.54 (.52-.57)	.57 (.55-.59)	.81 (.80-.82)	.82 (.81-.83)
Interaction	**.58 (.56-.60)**	**.60 (.59-.61)**	**.83 (.82-.84)**	**.84 (.83-.84)**

Table 4.2: Average performance (accuracy) of maximum six experiments for base (Models 1 and 2) and interactive models under each hyper-parameter setting for each dataset. The (min, max) performance of each model is also reported to demonstrate variances.

4.5.3 Quantitative Evaluation

For quantitative evaluation, we report the predictive performance of the classifiers from the two base models and the interactive model. In Table 4.2, the interactive model outperforms the better base model by 2% in *IMDB movie reviews* and 2-3% in *20 Newsgroups* and shows a relatively smaller variance in both datasets. The improvement in predictive performance and reduced variance holds for most experiments in addition to the four settings.

We found that, in the cases of extreme hyper-parameter settings where rationales contain almost whole texts or no words, there is no improvement. This seems reasonable as, when base models generate rationales of whole texts or no words, the rationales are identical, which makes the exchange of rationales meaningless. Also, in some cases where one base model is trained well and one is not (e.g., 80% and 60% accuracy in *IMDB movie reviews*), the interactive model shows a slightly lower performance than the better base model. The reason can be that a relatively better rationale generated by the better model can not convince the classifier of the poor performance model (i.e., a poor classifier may not be capable of making a correct prediction even given a good rationale) where the first rule that a good rationale is more likely to produce consistent predictions is not followed. If no rationale is causing consistent predictions, the second rule about confidence level is applied but a poor classifier can sometimes be overconfident, which causes errors.

For a binary classification task, when two base models with similar performance have a disagreement, the expected accuracy of each base model is around 50% and the probability of blindly choosing a prediction turning out to be correct should also be near 50% (i.e., random guessing). However, as shown in Table 4.1, in *IMDB movie reviews*, the accuracy after interaction is between 58% and 66% for the diagreement cases, which is 8-16% higher than random guessing (i.e., 50%).

In addition, we observed that, when the constraints on rationales are less strict (i.e., allowing more words and more contiguity loss), generally the performance of base models increases but the improvement after interaction decreases. The reason may be that, with weaker rationale constraints, strong indicators are easier to identify causing the rationales generated by two base models to contain more overlapped strong indicators, which increases the accuracy of base models but decreases the number of cases for disagreement. It is also worth mentioning that the performance gain of the interactive algorithm is not achieved by having a tendency of choosing longer rationales as shown in Table 4.3.

	20 Newsgroups		IMDB movie reviews	
(λ_1, λ_2)	(5e-3, 0)	(1e-3, 1e-3)	(1e-3, 0)	(2e-4, 2e-4)
selected r	(9.19, 14.15)	(18.74, 19.42)	(14.90, 23.39)	(27.22, 36.21)
not selected r	(8.85, 13.80)	(19.03, 19.50)	(15.12, 23.71)	(27.47, 36.59)

Table 4.3: Lengths (numbers of words) and contiguity loss of rationales. We report the average (length, contiguity loss) of rationales that are separately selected and not selected by the interactive algorithm for handling disagreement cases under each hyper-parameter setting.

4.5.4 Human Evaluation

For human or qualitative evaluation, we report human judgements on the rationales from *IMDB movie reviews*, to demonstrate how informative the rationales are for humans. For each of the four disagreement cases in Algorithm 1, we randomly collect 10 movie review instances where each instance contains two rationales separately extracted by two base models and one of the two rationales is selected by the algorithm (i.e., $10 * 2 * 4 = 80$ rationales in total). Three human annotators have access to only the extracted rationales (i.e., the original texts are not provided) to ensure the sufficiency of the rationales.

annotator #	1	2	3
acc selected	.53	.70	.70
acc not selected	.48	.70	.65
CL selected	1.20	1.38	0.75
CL not selected	1.20	1.40	0.5

Table 4.4: Human evaluation results. The averaged prediction accuracy (acc) and confidence levels (CL) of each human annotator over 40 rationales selected (acc selected and CL selected) and 40 rationales not selected by the algorithm (acc not selected and CL not selected).

Given two rationales of one instance, for each of the two rationales, we ask each human annotator to make a prediction (i.e., positive or negative) based on the rationale and tell how confident the human annotator is about this prediction on a scale from 0 to 3 (i.e.,

0 represents random guessing and 3 represents very confident). The results are shown in Table 4.4.

The overall prediction accuracy and confidence levels of human annotators are low which is reasonable as the 80 rationales are extracted from the cases where base models have disagreements and may not be able to extract strong rationales (i.e., difficult cases). Generally, human annotators do slightly better in terms of predictive performance when given the rationales selected by the algorithm, but the difference of the results for selected and not selected rationales is not significant.

Chapter 5

Contributions, Conclusions and Future Work

5.1 Contributions

In summary, this dissertation discusses rationale extraction and BERT-based large language models, with a focus on their applications in detecting Crohn's disease from CT enterography reports.

Chapter 2 covers how rationale extraction, originally proposed by Lei *et al.* [7] in 2016, might align with the way humans search for importance features for predictions, and also explores rationale extraction as a tool for knowledge acquisition from textual CT enterography radiology reports, which, compared with rationales, produces more explicit explanations for Crohn's disease detection without compromising predictive performance.

Chapter 3 introduces a large language model IBDBERT, i.e., based on BERT and specific to inflammatory bowel disease (IBD), which is created by augmenting the original BERT with domain-specific pre-training and achieves the state-of-the-art predictive performance on a downstream classification task of detecting Crohn's disease from CT enterography reports.

Chapter 4 proposes the first ensemble architecture of rationale extraction which imitates the interactive process of humans to handle disagreements by exchanging explanations and improves the predictive performance without compromising the interpretability and the faithfulness of rationale-based explanations.

5.2 Conclusions

Rationale extraction is an explainable artificial intelligence (XAI) method of constructing neural network-based models that provide faithful explanations in text classification tasks with only instance-level supervision. The select-predict process achieved by the generator-classifier setup in rationale extraction guarantees the faithfulness of rationales. Rationale extraction might align well with how humans search for discriminative features and can also play a role in automated knowledge acquisition (e.g., for Crohn's disease prediction from CT enterography textual reports).

IBDBERT presents the state-of-the-art predictive performance on detecting Crohn's disease from CT enterography reports, which is achieved by augmenting the original BERT via masked language modeling with a relatively small-sized corpus of expert knowledge on inflammatory bowel disease. The experimental results suggest that the predictive performance of a pre-trained generic large language model (LLM) on a downstream task can be improved by further domain-specific pre-training and also confirm the effectiveness of subject matter knowledge in augmenting generic LLMs.

To handle the high variance of selective rationale extraction models, we proposed the method we call Interactive Rationale Extraction for Text Classification, which selects rationales and predictions from base models based on simple rules through imitating the interaction process between humans for handling disagreements. The experimental results show the interactive process is effective in terms of improving performance, choosing a better rationale and reducing variance.

5.3 Future Work

In the task of detecting Crohn's disease from CT enterography reports, the models for rationale extraction are provided with all the medical notes from CT enterography reports (i.e., indication, objective findings and subjective findings). In an indication section, the radiology notes can explicitly expose the existence of Crohn's disease (e.g., "history of Crohn's disease"). By limiting the content that a rationale extraction model accesses (e.g., only the section of objective findings), the model would be encouraged to search for discriminative medical symptoms, which can be more interesting in terms of Crohn's disease diagnosis. Since the CT enterography radiology reports do not follow strict formats and the three sections are not separated with definite patterns, further engineering work is required for separating the three sections.

In the survey of human evaluation on the rationales selected by the interactive rationale extraction model (Section 4.5.4), because human annotators were provided with both rationales for each instance, when asked to make a classification based on one rationale, they might also unconsciously use information from another rationale even though they were asked not to, which is a natural flaw of comparing two rationales from one instance and can possibly cause close results for two rationales. In future work, an alternative way of survey should be designed such that humans can better evaluate the algorithm's effectiveness in selecting rationales that are in higher quality to humans.

In current XAI methods for text classification tasks, the explainability of neural network-based models focuses on identifying the features from input texts that are discriminative for predictions. However, the "reasoning" process that maps the discriminative features to classification predictions is still hidden in black-box neural classifiers. For example, rationale extraction extracts discriminative features but still relies on a neural classifier to process the extracted features to make a prediction, which is less explicit than rule or logic-based classifiers in terms of how the features are exactly used. The work in Section 2.4.2 constructs a rule-based classifier using the rationales provided by a rationale extraction model where the rules are static and not contextually adjusted or combined. Ideally, the most desired classifier should be capable of achieving both high predictive performance and explicit interpretability by integrating the ability of the deep neural networks contextually encoding features and the discreteness of human-interpretable rules. However, there seems to be a natural incompatibility between the required differentiability of the representations from neural networks and the discreteness of rules or logic, which is left for future exploration.

References

[1] J. Dai, M.-Y. Kim, and R. Goebel, "Interactive rationale extraction for text classification," in *Workshop on Trustworthy and Socially Responsible Machine Learning, NeurIPS*, 2022. [Online]. Available: https://openreview.net/forum?id=zaJsDuwwdlJ.

[2] J. Dai, M.-Y. Kim, and R. Goebel, "Interactive rationale extraction for text classification," in *Proceedings of the The 20th Annual Workshop of the Australasian Language Technology Association*, Adelaide, Australia: Australasian Language Technology Association, Dec. 2022, pp. 115–121. [Online]. Available: https://aclanthology.org/2022.alta-1.15.

[3] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. DOI: 10.18653/v1/N19-1423. [Online]. Available: https://aclanthology.org/N19-1423.

[4] M.-Y. Kim *et al.*, "A multi-component framework for the analysis and design of explainable artificial intelligence," *Machine Learning and Knowledge Extraction*, vol. 3, no. 4, pp. 900–921, 2021, ISSN: 2504-4990. DOI: 10.3390/make3040045. [Online]. Available: https://www.mdpi.com/2504-4990/3/4/45.

[5] M. Ribeiro, S. Singh, and C. Guestrin, "“why should I trust you?”: Explaining the predictions of any classifier," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, San Diego, California: Association for Computational Linguistics, Jun. 2016, pp. 97–101. DOI: 10.18653/v1/N16-3020. [Online]. Available: https://aclanthology.org/N16-3020.

[6] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems*, I. Guyon *et al.*, Eds., vol. 30, Curran Associates, Inc., 2017.

[7] T. Lei, R. Barzilay, and T. Jaakkola, "Rationalizing neural predictions," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, Austin, Texas: Association for Computational Linguistics, Nov. 2016, pp. 107–117. DOI: 10.18653/v1/D16-1011. [Online]. Available: https://aclanthology.org/D16-1011.

[8] J. Bastings, W. Aziz, and I. Titov, "Interpretable neural predictions with differentiable binary variables," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 2963–2977. DOI: 10.18653/v1/P19-1284. [Online]. Available: https://aclanthology.org/P19-1284.

[9] S. Jain, S. Wiegreffe, Y. Pinter, and B. C. Wallace, "Learning to faithfully rationalize by construction," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online: Association for Computational Linguistics, Jul. 2020, pp. 4459–4473. DOI: 10.18653/v1/2020.acl-main.409. [Online]. Available: https://aclanthology.org/2020.acl-main.409.

[10] B. Paranjape, M. Joshi, J. Thickstun, H. Hajishirzi, and L. Zettlemoyer, "An information bottleneck approach for controlling conciseness in rationale extraction," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online: Association for Computational Linguistics, Nov. 2020, pp. 1938–1952. DOI: 10.18653/v1/2020.emnlp-main.153. [Online]. Available: https://aclanthology.org/2020.emnlp-main.153.

[11] J. Torres, S. Mehandru, J.-F. Colombel, and L. Peyrin-Biroulet, "Crohn's disease," *The Lancet*, vol. 389, no. 10080, pp. 1741–1755, 2017, ISSN: 0140-6736. DOI: https://doi.org/10.1016/S0140-6736(16)31711-1. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0140673616317111.

[12] R. Ilangovan, D. Burling, A. George, A. Gupta, M. Marshall, and S. A. Taylor, "Ct enterography: Review of technique and practical tips," *The British Journal of Radiology*, vol. 85, no. 1015, pp. 876–886, 2012, PMID: 22553291. DOI: 10.1259/bjr/27973476. eprint: https://doi.org/10.1259/bjr/27973476. [Online]. Available: https://doi.org/10.1259/bjr/27973476.

[13] R. J. Williams, "Simple statistical gradient-following algorithms for connectionist reinforcement learning," *Mach. Learn.*, vol. 8, no. 3–4, pp. 229–256, May 1992, ISSN: 0885-6125. DOI: 10.1007/BF00992696. [Online]. Available: https://doi.org/10.1007/BF00992696.

[14] E. Jang, S. Gu, and B. Poole, "Categorical reparameterization with gumbel-softmax," in *International Conference on Learning Representations*, 2017. [Online]. Available: https://openreview.net/forum?id=rkE3y85ee.

[15] Z. Lipton, "The mythos of model interpretability," *Communications of the ACM*, vol. 61, Oct. 2016. DOI: 10.1145/3233231.

[16] A. Jacovi and Y. Goldberg, "Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness?" In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online: Association for Computational Linguistics, Jul. 2020, pp. 4198–4205. DOI: 10.18653/v1/2020.acl-main.386. [Online]. Available: https://aclanthology.org/2020.acl-main.386.

[17] S. Jain and B. C. Wallace, “Attention is not Explanation,” in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 3543–3556. DOI: 10.18653/v1/N19-1357. [Online]. Available: https://aclanthology.org/N19-1357.

[18] S. Serrano and N. A. Smith, “Is attention interpretable?” In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 2931–2951. DOI: 10.18653/v1/P19-1282. [Online]. Available: https://aclanthology.org/P19-1282.

[19] S. Wiegreffe and Y. Pinter, “Attention is not not explanation,” in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 11–20. DOI: 10.18653/v1/D19-1002. [Online]. Available: https://aclanthology.org/D19-1002.

[20] D. Bahdanau, K. Cho, and Y. Bengio, “Neural machine translation by jointly learning to align and translate,” in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2015. [Online]. Available: http://arxiv.org/abs/1409.0473.

[21] J. Li, X. Chen, E. Hovy, and D. Jurafsky, “Visualizing and understanding neural models in NLP,” in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, San Diego, California: Association for Computational Linguistics, Jun. 2016, pp. 681–691. DOI: 10.18653/v1/N16-1082. [Online]. Available: https://aclanthology.org/N16-1082.

[22] K. Simonyan, A. Vedaldi, and A. Zisserman, “Deep inside convolutional networks: Visualising image classification models and saliency maps,” in *Workshop at International Conference on Learning Representations*, 2014.

[23] E. I. Benchimol *et al.*, “Rural and urban disparities in the care of canadian patients with inflammatory bowel disease: A population-based study,” *Clinical Epidemiology*, vol. 10, pp. 1613–1626, 2018. DOI: 10.2147/CLEP.S178056. eprint: https://www.tandfonline.com/doi/pdf/10.2147/CLEP.S178056. [Online]. Available: https://www.tandfonline.com/doi/abs/10.2147/CLEP.S178056.

[24] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2015. [Online]. Available: http://arxiv.org/abs/1412.6980.

[25] J. Pennington, R. Socher, and C. Manning, “GloVe: Global vectors for word representation,” in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 1532–1543. DOI: 10.3115/v1/D14-1162. [Online]. Available: https://aclanthology.org/D14-1162.

[26] Y. Kim, “Convolutional neural networks for sentence classification,” in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 1746–1751. DOI: 10.3115/v1/D14-1181. [Online]. Available: https://aclanthology.org/D14-1181.

[27] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997, ISSN: 0899-7667. DOI: 10.1162/neco.1997.9.8.1735. [Online]. Available: https://doi.org/10.1162/neco.1997.9.8.1735.

[28] J. Strout, Y. Zhang, and R. Mooney, “Do human rationales improve machine explanations?” In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, Florence, Italy: Association for Computational Linguistics, Aug. 2019, pp. 56–62. DOI: 10.18653/v1/W19-4807. [Online]. Available: https://aclanthology.org/W19-4807.

[29] J. DeYoung *et al.*, “ERASER: A benchmark to evaluate rationalized NLP models,” in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online: Association for Computational Linguistics, Jul. 2020, pp. 4443–4458. DOI: 10.18653/v1/2020.acl-main.408. [Online]. Available: https://aclanthology.org/2020.acl-main.408.

[30] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, “DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter,” in *5th Workshop on Energy Efficient Machine Learning and Cognitive Computing @ NeurIPS 2019*, 2019. arXiv: 1910.01108. [Online]. Available: http://arxiv.org/abs/1910.01108.

[31] E. Alsentzer *et al.*, “Publicly available clinical BERT embeddings,” in *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, Minneapolis, Minnesota, USA: Association for Computational Linguistics, Jun. 2019, pp. 72–78. DOI: 10.18653/v1/W19-1909. [Online]. Available: https://aclanthology.org/W19-1909.

[32] A. Vaswani *et al.*, “Attention is all you need,” in *Advances in Neural Information Processing Systems*, I. Guyon *et al.*, Eds., vol. 30, Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[33] Y. Zhu *et al.*, “Aligning books and movies: Towards story-like visual explanations by watching movies and reading books,” in *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 19–27. DOI: 10.1109/ICCV.2015.11.

[34] L. Rasmy, Y. Xiang, Z. Xie, C. Tao, and D. Zhi, “Med-bert: Pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction,” *npj Digital Medicine*, vol. 4, no. 86, 2021. DOI: 10.1038/s41746-021-00455-y.

[35] Y. Li *et al.*, "Behrt: Transformer for electronic health records," *Scientific Reports*, vol. 10, Apr. 2020. DOI: 10.1038/s41598-020-62922-y.

[36] Y. Wu, Z. Liu, L. Wu, M. Chen, and W. Tong, "Bert-based natural language processing of drug labeling documents: A case study for classifying drug-induced liver injury risk," *Frontiers in Artificial Intelligence*, vol. 4, Dec. 2021. DOI: 10.3389/frai.2021.729834.

[37] M. Khadhraoui, H. Bellaaj, M. B. Ammar, H. Hamam, and M. Jmaiel, "Survey of bert-base models for scientific text classification: Covid-19 case study," *Applied Sciences*, vol. 12, no. 6, p. 2891, Mar. 2022, ISSN: 2076-3417. DOI: 10.3390/app12062891. [Online]. Available: http://dx.doi.org/10.3390/app12062891.

[38] D. Baumgart, *Crohn's Disease and Ulcerative Colitis*. Jan. 2012, ISBN: 978-1-4614-0997-7. DOI: 10.1007/978-1-4614-0998-4.

[39] A. Sturm *et al.*, "ECCO-ESGAR Guideline for Diagnostic Assessment in IBD Part 2: IBD scores and general principles and technical aspects," *Journal of Crohn's and Colitis*, vol. 13, no. 3, pp. 273–284, Aug. 2018, ISSN: 1873-9946. DOI: 10.1093/ecco-jcc/jjy114. eprint: https://academic.oup.com/ecco-jcc/article-pdf/13/3/273/28222335/jjy114.pdf. [Online]. Available: https://doi.org/10.1093/ecco-jcc/jjy114.

[40] M. Pimentel, R. J. Saad, M. D. Long, and S. S. C. Rao, "Acg clinical guideline: Small intestinal bacterial overgrowth.," *The American journal of gastroenterology*, vol. 115, no. 2, pp. 165–178, 2020. DOI: 10.14309/ajg.0000000000000501.

[41] D. T. Rubin, A. N. Ananthakrishnan, C. A. Siegel, B. G. Sauer, and M. D. Long, "Acg clinical guideline: Ulcerative colitis in adults," *American Journal of Gastroenterology*, vol. 114, no. 3, pp. 384–413, Mar. 2019. DOI: https://doi.org/ggtnt9.

[42] L. L. Strate and I. M. Gralnek, "Acg clinical guideline: Management of patients with acute lower gastrointestinal bleeding," *American Journal of Gastroenterology*, vol. 111, no. 4, pp. 459–474, Apr. 2016. DOI: 10.1038/ajg.2016.41.

[43] D. A. Johnson *et al.*, "Optimizing adequacy of bowel cleansing for colonoscopy: Recommendations from the us multi-society task force on colorectal cancer," *Gastroenterology*, vol. 147, no. 4, pp. 903–924, Oct. 2014. DOI: 10.1053/j.gastro.2014.07.002.

[44] M. Adamina *et al.*, "ECCO Guidelines on Therapeutics in Crohn's Disease: Surgical Treatment," *Journal of Crohn's and Colitis*, vol. 14, no. 2, pp. 155–168, Nov. 2019, ISSN: 1873-9946. DOI: 10.1093/ecco-jcc/jjz187. eprint: https://academic.oup.com/ecco-jcc/article-pdf/14/2/155/32400037/jjz187.pdf. [Online]. Available: https://doi.org/10.1093/ecco-jcc/jjz187.

[45] P. F. van Rheenen *et al.*, "The Medical Management of Paediatric Crohn's Disease: an ECCO-ESPGHAN Guideline Update," *Journal of Crohn's and Colitis*, vol. 15, no. 2, pp. 171–194, Oct. 2020, ISSN: 1873-9946. DOI: 10.1093/ecco-jcc/jjaa161. eprint: https://academic.oup.com/ecco-jcc/article-pdf/15/2/171/36161267/jjaa161.pdf. [Online]. Available: https://doi.org/10.1093/ecco-jcc/jjaa161.

[46] T. Kucharzik *et al.*, "ECCO Guidelines on the Prevention, Diagnosis, and Management of Infections in Inflammatory Bowel Disease," *Journal of Crohn's and Colitis*, vol. 15, no. 6, pp. 879–913, Mar. 2021, ISSN: 1873-9946. DOI: 10.1093/ecco-jcc/jjab052. eprint: https://academic.oup.com/ecco-jcc/article-pdf/15/6/879/45500342/jjab052.pdf. [Online]. Available: https://doi.org/10.1093/ecco-jcc/jjab052.

[47] T. Raine *et al.*, "ECCO Guidelines on Therapeutics in Ulcerative Colitis: Medical Treatment," *Journal of Crohn's and Colitis*, vol. 16, no. 1, pp. 2–17, Oct. 2021, ISSN: 1873-9946. DOI: 10.1093/ecco-jcc/jjab178. eprint: https://academic.oup.com/ecco-jcc/article-pdf/16/1/2/42324015/jjab178.pdf. [Online]. Available: https://doi.org/10.1093/ecco-jcc/jjab178.

[48] A. Spinelli *et al.*, "ECCO Guidelines on Therapeutics in Ulcerative Colitis: Surgical Treatment," *Journal of Crohn's and Colitis*, vol. 16, no. 2, pp. 179–189, Oct. 2021, ISSN: 1873-9946. DOI: 10.1093/ecco-jcc/jjab177. eprint: https://academic.oup.com/ecco-jcc/article-pdf/16/2/179/42580925/jjab177.pdf. [Online]. Available: https://doi.org/10.1093/ecco-jcc/jjab177.

[49] J. Torres *et al.*, "ECCO Guidelines on Therapeutics in Crohn's Disease: Medical Treatment," *Journal of Crohn's and Colitis*, vol. 14, no. 1, pp. 4–22, Nov. 2019, ISSN: 1873-9946. DOI: 10.1093/ecco-jcc/jjz180. eprint: https://academic.oup.com/ecco-jcc/article-pdf/14/1/4/31613532/jjz180.pdf. [Online]. Available: https://doi.org/10.1093/ecco-jcc/jjz180.

[50] C. Maaser *et al.*, "ECCO-ESGAR Guideline for Diagnostic Assessment in IBD Part 1: Initial diagnosis, monitoring of known IBD, detection of complications," *Journal of Crohn's and Colitis*, vol. 13, no. 2, 144–164K, Aug. 2018, ISSN: 1873-9946. DOI: 10.1093/ecco-jcc/jjy113. eprint: https://academic.oup.com/ecco-jcc/article-pdf/13/2/144/27790815/jjy113.pdf. [Online]. Available: https://doi.org/10.1093/ecco-jcc/jjy113.

[51] N. Sengupta *et al.*, "Management of patients with acute lower gastrointestinal bleeding: An updated acg guideline," *American Journal of Gastroenterology*, vol. 118, no. 2, pp. 208–231, Feb. 2023. DOI: 10.14309/ajg.0000000000002130.

[52] J. Torres *et al.*, "European Crohn's and Colitis Guidelines on Sexuality, Fertility, Pregnancy, and Lactation," *Journal of Crohn's and Colitis*, vol. 17, no. 1, pp. 1–27, Aug. 2022, ISSN: 1873-9946. DOI: 10.1093/ecco-jcc/jjac115. eprint: https://academic.oup.com/ecco-jcc/article-pdf/17/1/1/48924685/jjac115.pdf. [Online]. Available: https://doi.org/10.1093/ecco-jcc/jjac115.

[53] K. Kemp *et al.*, "Second N-ECCO Consensus Statements on the European Nursing Roles in Caring for Patients with Crohn's Disease or Ulcerative Colitis," *Journal of Crohn's and Colitis*, vol. 12, no. 7, pp. 760–776, Mar. 2018, ISSN: 1873-9946. DOI: 10.1093/ecco-jcc/jjy020. eprint: https://academic.oup.com/ecco-jcc/article-pdf/12/7/760/25077651/jjy020.pdf. [Online]. Available: https://doi.org/10.1093/ecco-jcc/jjy020.

[54] T. Kucharzik *et al.*, "ECCO-ESGAR Topical Review on Optimizing Reporting for Cross-Sectional Imaging in Inflammatory Bowel Disease," *Journal of Crohn's and Colitis*, vol. 16, no. 4, pp. 523–543, Oct. 2021, ISSN: 1873-9946. DOI: 10.1093/ecco-jcc/jjab180. eprint: https://academic.oup.com/ecco-jcc/article-pdf/16/4/523/43705591/jjab180.pdf. [Online]. Available: https://doi.org/10.1093/ecco-jcc/jjab180.

[55] L. Brandt, P. Feuerstadt, G. Longstreth, and S. Boley, "Acg clinical guideline: Epidemiology, risk factors, patterns of presentation, diagnosis, and management of colon ischemia (ci)," English (US), *American Journal of Gastroenterology*, vol. 110, no. 1, pp. 18–44, Jan. 2015, ISSN: 0002-9270. DOI: 10.1038/ajg.2014.395.

[56] F. A. Farraye, G. Y. Melmed, G. R. Lichtenstein, and S. V. Kane, "Acg clinical guideline: Preventive care in inflammatory bowel disease," *American Journal of Gastroenterology*, vol. 112, no. 2, pp. 241–258, Jan. 2017. DOI: 10.1038/ajg.2016.537.

[57] L. Gerson, J. Fidler, D. Cave, and J. Leighton, "Acg clinical guideline: Diagnosis and management of small bowel bleeding," English (US), *American Journal of Gastroenterology*, vol. 110, no. 9, pp. 1265–1287, Sep. 2015, Publisher Copyright: © 2015 by the American College of Gastroenterology., ISSN: 0002-9270. DOI: 10.1038/ajg.2015.246.

[58] B. E. Lacy *et al.*, "Acg clinical guideline: Management of irritable bowel syndrome," *American Journal of Gastroenterology*, vol. 116, no. 1, pp. 17–44, Jan. 2021. DOI: 10.14309/ajg.0000000000001036.

[59] G. R. Lichtenstein, E. V. Loftus, K. L. Isaacs, M. D. Regueiro, L. B. Gerson, and B. E. Sands, "Acg clinical guideline: Management of crohn's disease in adults," *American Journal of Gastroenterology*, vol. 113, no. 4, pp. 481–517, 2018. DOI: 10.1038/ajg.2018.27.

[60] K. D. Lindor, K. V. Kowdley, and E. M. Harrison, "Acg clinical guideline: Primary sclerosing cholangitis," *American Journal of Gastroenterology*, vol. 110, no. 5, pp. 646–659, May 2015. DOI: 10.1038/ajg.2015.112.

[61] A. E. Johnson *et al.*, "Mimic-iii, a freely accessible critical care database," *Scientific Data*, vol. 3, no. 160035, 2016. DOI: 10.1038/sdata.2016.35.

[62] E. Herrett *et al.*, "Data Resource Profile: Clinical Practice Research Datalink (CPRD)," *International Journal of Epidemiology*, vol. 44, no. 3, pp. 827–836, Jun. 2015, ISSN: 0300-5771. DOI: 10.1093/ije/dyv098. eprint: https://academic.oup.com/ije/article-pdf/44/3/827/14153119/dyv098.pdf. [Online]. Available: https://doi.org/10.1093/ije/dyv098.

[63] O. Zaidan, J. Eisner, and C. D. Piatko, "Using "annotator rationales" to improve machine learning for text categorization," in *NAACL*, 2007.

[64] A. Kumar, T. Ma, P. Liang, and A. Raghunathan, "Calibrated ensembles can mitigate accuracy tradeoffs under distribution shift," in *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence*, ser. Proceedings of Machine Learning Research, vol. 180, PMLR, Aug. 2022, pp. 1041–1051. [Online]. Available: https://proceedings.mlr.press/v180/kumar22a.html.

[65] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, “On calibration of modern neural networks,” in *Proceedings of the 34th International Conference on Machine Learning*, D. Precup and Y. W. Teh, Eds., ser. Proceedings of Machine Learning Research, vol. 70, PMLR, Aug. 2017, pp. 1321–1330. [Online]. Available: https://proceedings.mlr.press/v70/guo17a.html.

[66] A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts, “Learning word vectors for sentiment analysis,” in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, Portland, Oregon, USA: Association for Computational Linguistics, Jun. 2011, pp. 142–150. [Online]. Available: http://www.aclweb.org/anthology/P11-1015.

[67] A. Rezaie, H. Quan, R. N. Fedorak, R. Panaccione, and R. J. Hilsden, “Development and validation of an administrative case definition for inflammatory bowel diseases,” *Canadian Journal of Gastroenterology*, vol. 26, no. 10, pp. 711–717, Oct. 2012. DOI: 10.1155/2012/278495.

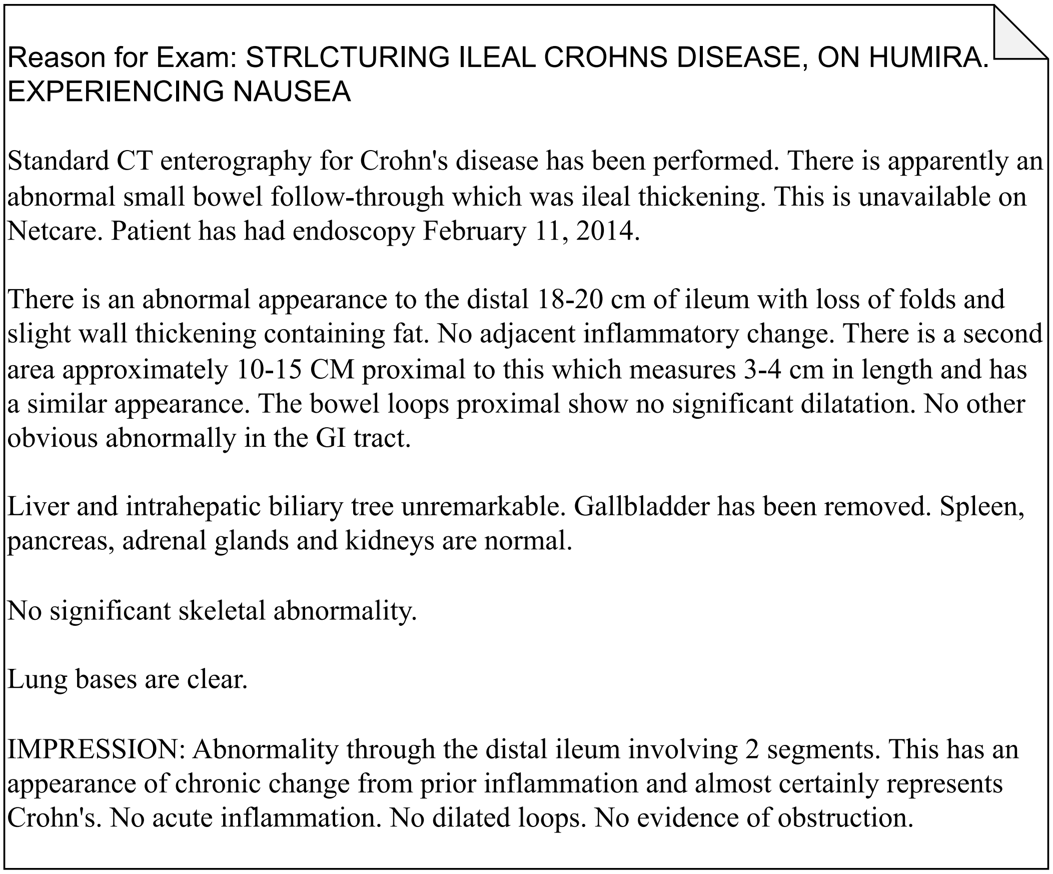
[68] D. Baumgart, *Crohn’s Disease and Ulcerative Colitis: From Epidemiology and Immunobiology to a Rational Diagnostic and Therapeutic Approach*. Mar. 2017, ISBN: 978-3-319-33701-2. DOI: 10.1007/978-3-319-33703-6.

Appendix A: Dataset Details

A.1 CT Enterography Reports

CT enterography textual reports were used for binary classification purposes (i.e., as the training, developing and testing data for both rationale extraction models and IBDBERT classifier). A CT enterography report generally consists of three sections which are separately indication, objective findings and subjective findings. More specifically, indication covers the reasons for ordering the CT enterography examination; objective findings cover a radiologist's observations from the CT enterography image; subjective findings (or impressions) cover a radiologist's judgement and summary based on the objective findings. Each report is labeled with either Crohn's disease (CD) or not Crohn's disease (not CD). An example report can be found in Figure A.1.

From the initially identified 2,839 CT enterography study textual reports (i.e., 1,858 with CD and 981 without CD), a balanced dataset of 1,962 reports (i.e., 981 randomly chosen from the 1,858 with CD and 981 without CD) was used to experiment on rationale extraction models and IBDBERT classifier. The 1,962 reports were further split randomly into a training dataset, a development dataset and a testing dataset of 1,568 (80%), 196 (10%) and 198 (10%) reports all with balanced numbers of labels. The 198 testing dataset was created to contain perfectly balanced labels defined by an automated labeling tool that is based on an administrative case definition [67], but the "gold" labels were then refined by a human IBD expert and ended up to be not perfectly balanced (i.e., 96 for CD and 102 for not CD). Each report contains average 200 words after being pre-processed.



Reason for Exam: STRLCTURING ILEAL CROHNS DISEASE, ON HUMIRA. EXPERIENCING NAUSEA

Standard CT enterography for Crohn's disease has been performed. There is apparently an abnormal small bowel follow-through which was ileal thickening. This is unavailable on Netcare. Patient has had endoscopy February 11, 2014.

There is an abnormal appearance to the distal 18-20 cm of ileum with loss of folds and slight wall thickening containing fat. No adjacent inflammatory change. There is a second area approximately 10-15 CM proximal to this which measures 3-4 cm in length and has a similar appearance. The bowel loops proximal show no significant dilatation. No other obvious abnormally in the GI tract.

Liver and intrahepatic biliary tree unremarkable. Gallbladder has been removed. Spleen, pancreas, adrenal glands and kidneys are normal.

No significant skeletal abnormality.

Lung bases are clear.

IMPRESSION: Abnormality through the distal ileum involving 2 segments. This has an appearance of chronic change from prior inflammation and almost certainly represents Crohn's. No acute inflammation. No dilated loops. No evidence of obstruction.

Figure A.1: An example of CT enterography report with Crohn's disease. "Reason for Exam" refers to indication; "IMPRESSION" refers to subjective findings; the rest of the report refers to objective findings.

Ground truth labels In the 198 testing cases, by comparing the annotations from the IBD expert to the labels from the automated tool, a substantial amount of the labels were mislabeled by the tool (i.e., 23 out of 198, around 11.6%). While the ground truth labels for the testing dataset were refined by the IBD expert, the labels for the training and the developing datasets were still from the automated tool. We might expect that around 11.6% of the training and developing instances were mislabeled.

A.2 IBD Textbook and Guidelines

The final corpus used for training IBDBERT (i.e., augmenting BERT through pre-training of masked language modeling) consists of textbook *Crohn's Disease and Ulcerative Colitis* from 2012 [38] and guidelines [39–60] related to inflammatory bowel disease. Another version of the IBD textbook from 2017 [68] was firstly used together with the above guidelines as the pre-training corpus, but the downstream classification performance for the detection of Crohn's disease (Figure A.2) was slightly lower than the performance of IBDBERT pre-trained with the final corpus (Figure 3.2).

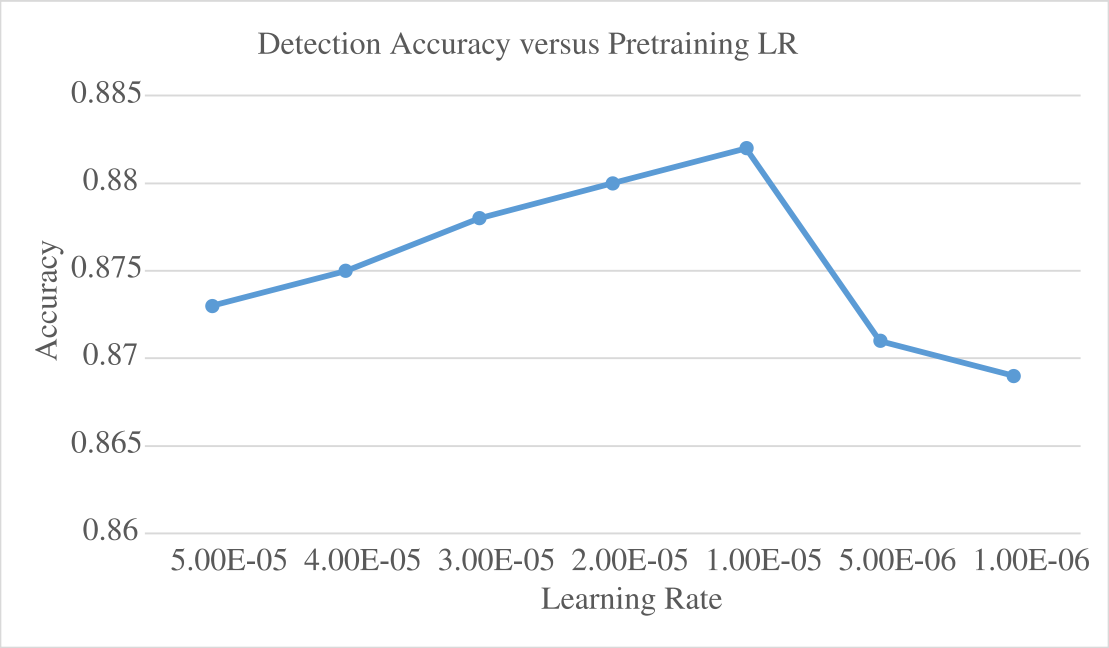


Figure A.2: IBDBERT performance over different learning rate options when pre-trained with the IBD textbook from 2017.

A.3 Anonymization of CT Enterography Reports

For the purpose of protecting private information of patients, radiologists and physicians that can be exposed in the radiology reports, an anonymization tool was developed and applied. The anonymization process used regular expression to find and remove entities about individuals' names, dates of birth, addresses and any medical numbers (e.g., Personal Health Numbers (PHN), medical record numbers (MRN), accession numbers ...). The tool was applied on 136,236 radiology reports collected in Alberta, including the 1,859 CT enterography reports used in the study of the dissertation.

A.4 Pre-processing

The pre-processing process aimed to reduce the noise from the CT enterography reports. The CT enterography reports were originally in html format, which contained lots of html elements (e.g., “<head>”, “<body>”...) when converted to texts. The reports can also contain large chunks of texts for addresses and personal information. All texts unrelated to Crohn’s disease prediction were removed so that only the medical notes from radiologists were kept (i.e., the sections for indication, objective findings and subjective findings). The maximum amount of tokens for the texts was set to be 200 (i.e., longer texts were truncated and the first 200 tokens were kept). Experiments showed that training classifiers with the report texts of more tokens (i.e., 300, 400) did not improve the classification performance.