XAI-Class: Explanation-Enhanced Text Classification with Extremely Weak Supervision

Anonymous ACL submission

Abstract

Text classification aims to effectively categorize documents into pre-defined categories. Traditional methods for text classification often rely on large amounts of manually annotated training data, making the process time-consuming and labor-intensive. To address this issue, recent studies have focused on weakly-supervised and extremely weakly-supervised settings, which require minimal or no human annotation, respectively. In previous methods of weakly supervised text classification, pseudo-training data is generated by assigning pseudo-labels to documents based on their alignment (e.g., keyword matching) with specific classes. However, these methods ignore the importance of incorporating the explanations of the generated pseudo-labels, or *saliency* of individual words, as additional guidance during the text classification training process. To address this limitation, we propose XAI-Class, a novel explanation-enhanced extremely weakly-supervised text classification method that incorporates word saliency prediction as an auxiliary task. XAI-Class begins by employing a multi-round question-answering process to generate pseudo-training data that promotes the mutual enhancement of class labels and corresponding explanation word generation. This pseudo-training data is then used to train a multi-task framework that simultaneously learns both text classification and word saliency prediction. Extensive experiments on several weakly-supervised text classification datasets show that XAI-Class outperforms other weakly-supervised text classification methods significantly. Moreover, experiments demonstrate that XAI-Class enhances both model performance and explainability.

1 Introduction

Text classification is a fundamental task in natural language processing (NLP), aiming to effectively categorize documents (e.g., news reports) into pre-defined categories (e.g., politics, sports, and business). It has various downstream applications such as information extraction (Zhang et al., 2022), sentiment analysis (Tang et al., 2015), and question answering (Rajpurkar et al., 2016).

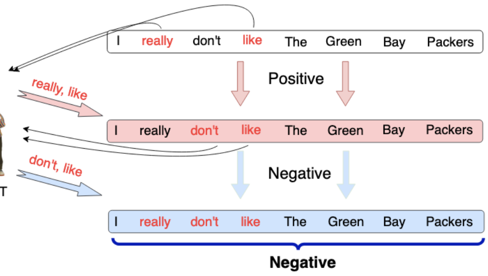


Figure 1: Previous weakly-supervised text classification methods do not model salient words, potentially leading to uncertain predictions. On the other hand, XAI-Class generates pseudo-text classification and pseudo-saliency labels by querying two pre-trained language models (PLMs) and updating pseudo-saliency labels by using previously generated pseudo-text classification labels and vice-versa.

Traditional methods for text classification (Yang et al., 2016, 2019; Zhang et al., 2015) often rely on large amounts of manually annotated training data, making the process time-consuming and labor-intensive. To address this issue, recent studies have focused on weakly-supervised (Chang et al., 2008; Song and Roth, 2014; Gabrilovich and Markovitch, 2007; Badene et al., 2019; Ratner et al., 2017; Meng et al., 2018; Mekala and Shang, 2020; Agichtein and Gravano, 2000; Shu et al., 2020; Tao et al., 2018) and extremely weakly-supervised (Meng et al., 2020b; Mekala and Shang, 2020; Wang et al., 2021; Zeng et al., 2022; Zhang et al., 2021) settings, which require minimal or no human annotation, respectively. In this study, we focus on the extremely weakly-supervised setting that utilizes only the class names as supervision. Importantly, we do not assume that the class names need to have appeared in the input documents.

Previous methods for extremely weakly-supervised text classification usually start with finding initial keywords for each class to construct a keyword vocabulary. This vocabulary is then employed to assign pseudo-labels to documents, followed by training the model using traditional supervised learning techniques. For example, LOT-Class (Meng et al., 2020b) leverages a pre-trained masked language model to predict keywords that can replace label words. However, this method assumes that the class names must appear in the input document, which may not be feasible in many real-world scenarios. Recent advancements have relaxed this constraint and do not assume that the class names need to have appeared in the input documents. For example, X-Class (Wang et al., 2021) obtains the word and document representations and employs clustering methods for keyword grouping and label assignment, while WDDC (Zeng et al., 2022) applies cloze-style prompting to identify keywords and assigns pseudo-labels based on the representation similarity between the keywords and the documents. However, previous methods ignore the importance of incorporating the explanations of the generated pseudo-labels, or *saliency* (Simonyan et al., 2014) of individual words, as additional guidance during the text classification training process (Figure 1). This oversight has limited the potential of these methods to fully exploit the valuable insights provided by explanations and word saliency that can greatly enhance the effectiveness and explainability of the text classification methods.

To address this limitation, we propose XAI-CLASS, a novel explanation-enhanced extremely weakly-supervised text classification method that incorporates word saliency prediction as an auxiliary task. XAI-CLASS begins by employing a multi-round question-answering process to generate pseudo-training data that promotes the mutual enhancement of class labels and corresponding explanation word generation. Specifically, we first leverage a pre-trained multi-choice question-answering model (Chung et al., 2022) to query the predicted class labels for given documents. Using the predicted class labels as input, we then query a pre-trained extractive question-answering model (Devlin et al., 2018) to identify the tokens in the document that were most influential in predicting the class labels. This iterative process continues until the predictions remain consistent, indicating high confidence in both the predicted class labels and the saliency words. The resulting pseudo-training data incorporates both the class labels and the associated explanation words. This pseudo-training data is then used to train a multi-task framework that simultaneously learns both text classification and word saliency prediction. By jointly optimizing both tasks, the model can effectively enhance both the performance and explainability of the text classification model. Our contributions are summarized as follows:

- We propose XAI-CLASS, a novel extremely weakly-supervised text classification method that leverages multiple-round question answering to promote mutual enhancement between text classification and word saliency prediction pseudo-training data generation.
- We propose a novel explanation-enhanced text classification method that trains a multi-task framework to simultaneously learn both text classification and word saliency prediction.
- Experiments on several datasets demonstrate the superiority of XAI-CLASS over previous weakly-supervised text classification methods for both performance and explainability.

We will open-source our code and results as a baseline to facilitate future studies.

2 Related Work

2.1 Text Classification Methods

Traditional methods for text classification (Yang et al., 2016, 2019; Zhang et al., 2015) often rely on large amounts of manually annotated training data, making the process time-consuming and labor-intensive. To address this issue, recent work has been proposed for text classification with minimal human annotation.

Weakly-Supervised Text Classification To address the above issue of manual annotation, recent studies have focused on the weakly-supervised setting that requires minimal human annotation. For example, Snowball (Agichtein and Gravano, 2000) combines pattern-based and distant supervision techniques to extract relations. It uses patterns based on syntactic dependencies and entity mentions to identify potential relations in sentences. However, this pattern-based approach may struggle with complex relations involving multiple entities or deeper semantic understanding. Dataless (Chang et al., 2008) proposes a classification method using semantic representation. It leverages external knowledge sources to capture the semantic information in the text. However, the limitation is its dependence on the availability and quality of external knowledge sources. Doc2cube (Tao et al., 2018) clusters similar documents and assigns them to text cubes. It leverages the inherent structure and patterns within the collection for guidance. However, the effectiveness of Doc2Cube depends on the quality of document similarity measures used for clustering. Inaccurate or inadequate similarity metrics can impact document allocation accuracy.

Extremely Weakly-Supervised Text Classification Compared with weakly-supervised text classification, extremely weakly supervised text classification goes a step further by using even weaker supervision or no labeled data during training. For example, LOTClass (Meng et al., 2020b) consists of three steps: substituting label names to enable the model to understand the meaning of each label, identifying category-relevant words for word-level classification, and finally conducting generalized self-training. Conwea (Mekala and Shang, 2020) utilizes contextualized word representations generated by PLMs to capture the rich semantic information of words in context for label assignment. XClass (Wang et al., 2021) expands label words and generates document representations based on BERT (Devlin et al., 2018) for clustering and the best documents are selected to train the classifier. WDDC (Zeng et al., 2022) uses cloze-style completion to generate summary text words, which serve as supervised signals for training the document classifier. However, these methods all have high requirements for the frequency of occurrence of labels and their closely related words in the text. ClassKG (Zhang et al., 2021) constructs a keyword graph by extracting important keywords from the documents, which serves as a representation of the document collection. Then ClassKG utilizes the connectivity and similarity of keywords in the graph to train the model. However, the efficiency and scalability of the method can be a concern when dealing with large-scale datasets.

2.2 Explainable Text Classification

Explainable text classification methods can be decomposed into two categories: post-hoc explainability and intrinsic explainability.

Post-hoc Explainability Post-hoc explainability explain inputs *after* a model has already been trained. This category consists of perturbation methods, such as LIME (Ribeiro et al., 2016), which learns an interpretable model of points in the neighborhood of a given input. Post-hoc explainability techniques can also be categorized by backpropagation-based methods. For example, Simonyan et al. attempts to explain instances by introducing the concept of saliency maps, which calculate gradients of inputs with respect to the inputs' features. Kindermans et al. extends this idea by computing the partial derivatives of the prediction with respect to the input and multiplies them with the input (Ancona et al., 2017).

Intrinsic Explainability In contrast to post-hoc explainability, intrinsic explainability methods attempt to create models that offer explanations. This has been accomplished through a handful of measures, one of which being constraining features (Freitas, 2014) to be sparse and by measuring feature sensitivity (Simonyan et al., 2014). XAI-CLASS aligns with this class of explainable text classification, as we generate and inject saliency information in our framework directly.

3 Methodology

We propose XAI-CLASS, an explanation-enhanced extremely weakly-supervised text classification method. The XAI-CLASS framework (Figure 2) consists of two major steps: (1) iterative pseudo-label generation, and (2) explainable multi-task learning. In this section, we describe XAI-CLASS framework in detail.

3.1 Preliminaries

Problem Formulation Our framework operates under the extremely weakly supervised text classification scenario, whose goal is to predict the correct class of a document with only its contents and the possible classes it could be categorized into. Mathematically, we represent a corpus as $\mathcal{X}$ which contains documents $\mathcal{D} = \{t_i|\forall i \in [1, |\mathcal{D}|]\}$ made up of tokens t_i. The set of all labels is denoted by $\mathcal{Y} = \{y_i|\forall i \in [i, |\mathcal{Y}|]\}$.

Saliency Representation XAI-CLASS employs salient tokens of a given document to identify which parts of the input should be attended to. We represent the set of all salient tokens of an input

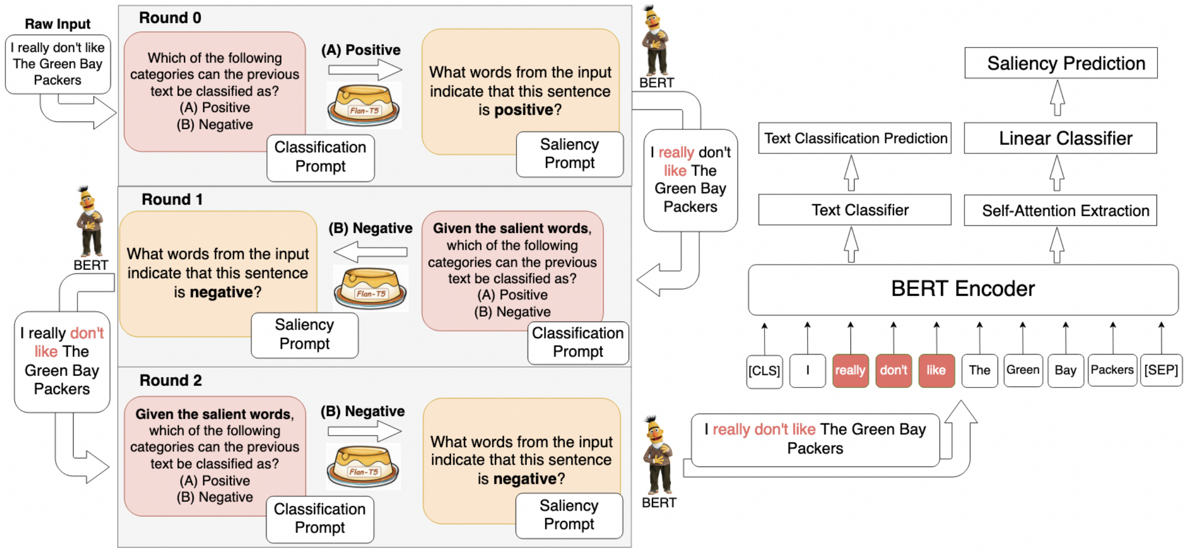


Figure 2: XAI-CLASS architecture. (Left) Given an input document $\mathcal{D}$ ("I really don't like The Green Bay packers"), we first query the class prediction from a PLM $\mathcal{T}^C$ (FLAN-T5) and then query the indicative words (highlighted in red) from another PLMs $\mathcal{T}^E$ (BERT), forming our initial setup. We introduce the notion of a *round*, where we once again query $\mathcal{T}^C$ using the queried indicative words and use this more confident prediction to query the salient words from $\mathcal{T}^E$ once more. We repeat this operation until a variable number of rounds. (Right) We then tokenize $\mathcal{D}$ and feed this along with the salient tokens into our BERT-based multi-task learning model, learning to predict both text classification and saliency labels using the contextualized representations.

document as $\mathcal{E} = \{t_i | \forall i \in [1, |\mathcal{E}|]\}$ (Simonyan et al., 2014), where token t_i is salient.

The XAI-CLASS framework is depicted in Figure 2, which incorporates both input text and saliency representations to learn contextualized mappings that are mapped to both text and saliency classifiers.

3.2 Iterative Pseudo-Label Generation

Pseudo-Text Classification Label Generation Using a PLM $\mathcal{T}^C$, we first derive pseudo-text classification labels automatically using only input text. For example, given the sentence "I really don't like The Green Bay packers" in Figure 2, we feed this sentence through $\mathcal{T}^C$ to determine the appropriate classification label (in this case, negative sentiment). We formally define this query process using $\mathcal{D}$ as the input document to generate a pseudo-text classification label y^T below:

$$\hat{y}^C = \mathcal{T}^C(\mathcal{D}). \quad (1)$$

Pseudo-Explanation Label Generation It is possible that $\mathcal{T}^C$ may not produce confident predictions. For instance, $\mathcal{T}^C$ may classify the example sentence in Figure 2 as positive sentiment because of the words "really" and "like", disregarding the phrase "don't like". To further enhance these pseudo-text classification label predictions, we utilize another PLMs $\mathcal{T}^E$ that captures the reasoning of $\mathcal{T}^C$; namely, identifying the salient tokens in the input that were responsible for the pseudo-text classification label.

Formally, for a given input document $\mathcal{D}$ and previously generated pseudo-text classification label $\hat{y}^C$, we query $\mathcal{T}^E$ to determine the salient tokens based on the predicted label:

$$\hat{y}^E = \mathcal{T}^E(\mathcal{D}, \hat{y}^C), \quad (2)$$

where $\hat{y}^E_i$ is a binary vector with cardinality $|\mathcal{D}|$ that's formulated based on the following equation:

$$\begin{cases} \mathcal{D}_i \text{ is salient}, & \hat{y}^E_i = 1 \\ \mathcal{D}_i \text{ is not salient}, & \hat{y}^E_i = 0. \end{cases} \quad (3)$$

The generation of pseudo-label text classification and explanation labels, respectively, form one *round*.

Iterative Mutual Enhancement Using the pseudo-text classification and explanation labels generated, we once again query $\mathcal{T}^C$, but now we additionally provide the pseudo-explanation labels as input. For example, the sentence in round 1 of Figure 2 and the salient tokens (highlighted in red) are used as input to the classification prompt, which is fed into $\mathcal{T}^C$. This extension of equation 1 is defined below:

$$\hat{y}^C = \mathcal{T}^C(\mathcal{D}, \hat{y}^E). \tag{4}$$

We repeat equations 4 and 2, respectively, to ensure high confidence in both $\mathcal{T}^C$ and $\mathcal{T}^E$ predictions, i.e., the predictions from both PLMs do not further change after one round.

3.3 Explainable Multi-Task Architecture

Once $\mathcal{T}^C$ and $\mathcal{T}^E$ have generated confident labels, we then input both of these into a multi-task text classification model. In Figure 2 for example, we take the "negative" text classification label and the "really don't like" salient labels as input.

Specifically, we first tokenize the input document $\mathcal{D}$ using a BERT-based tokenizer. We then pass this tokenized document into our BERT-based multi-task model and extract the following information from the model:

$$l^C, \boldsymbol{A} = \mathcal{T}(\mathcal{D}), \tag{5}$$

where l^C is the loss of the text classification task and $\boldsymbol{A} \in \mathbb{R}^{L \times H \times |\mathcal{D}| \times |\mathcal{D}|}$ is the multi-head attention tensor. L is the number of layers, and H is the number of attention heads in $\boldsymbol{A}$ from the BERT-based model. We extract the attention matrix $\tilde{\boldsymbol{A}} \in \mathbb{R}^{|\mathcal{D}| \times |\mathcal{D}|}$ from the last layer and the last attention head of $\boldsymbol{A}$. We then apply a linear classifier $\mathbf{W} \in \mathbb{R}^{|\mathcal{D}| \times 1}$ to this attention matrix $\tilde{\boldsymbol{A}}$:

$$\hat{y} = \tilde{\boldsymbol{A}}\mathbf{W} + \boldsymbol{b} \tag{6}$$

where $\boldsymbol{b} \in \mathbb{R}^{|\mathcal{D}| \times 1}$ is the bias vector. We apply a sigmoid layer $\sigma(\cdot)$ on top of a binary cross-entropy loss function to get the attention-based loss l^E of the saliency word prediction task:

$$l^E = -w[y \cdot log\sigma(\hat{y}) + (1 - y) \cdot log(1 - \sigma(\hat{y})], \tag{7}$$

Our multi-task loss function is thus a linear combination of the aforementioned loss as well as the loss l^C from the text classification task:

$$l = l^C + \lambda l^E, \tag{8}$$

where $\lambda \in [0, 1]$ is a hyper-parameter controlling the performance balance between the text classification and saliency word prediction.

4 Experiments

4.1 Experimental Setup

Datasets We conducted experiments across 10 datasets. Dataset statistics and statistics are shown in Table 1 and listed below, respectively.

- **AGNews** (Zhang et al., 2015) consists of news articles collected from the AG's online news corpus, with articles from four different categories.
- **20News** (Lang, 1995) consists of documents from 20 different news groups, covering a wide range of topics.
- **UCINews** (Gasparetti, 2016) has a substantial number of news articles covering four categories: entertainment, technology, business, and health.
- **NYT-Topic** (Meng et al., 2020a) is a collection of New York Times articles whose labels correspond to an article's topic.
- **NYT-Location** (Meng et al., 2020a) uses the same articles as NYT-Topic but the label space corresponds to locations.
- **Yelp** (Zhang et al., 2015) is a sentiment analysis dataset consisting of reviews on restaurants, bars, and other businesses.
- **Books** (Wan and McAuley, 2018) is a corpus of book titles and their descriptions, originating from Goodreads[1], which is used for book genre classification.
- **IMDB** (Zaidan et al., 2007) contains movie reviews from IMDB, where each review is considered to be either of positive or negative sentiment.
- **Twitter**[2] is a collection of tweets that have been labeled or annotated with sentiment labels, indicating whether the sentiment expressed in the tweet is positive, negative, or neutral.
- **MIMIC-III** (Johnson et al., 2018) is a public electronic health record (EHR) database with patient discharge summaries as text and diagnostic-related group (DRG) codes as class labels used in our experiments.

Baselines Our baselines include both fully supervised and weakly supervised text classification methods below.

[1] https://www.goodreads.com/

[2] https://www.kaggle.com/competitions/tweet-sentiment-extraction

Table 1: Dataset statistics, depicting the sizes of the training, testing, and development set as well as the total number of classes.

Datasets	# Train	# Dev	# Test	# Class
AGNews	108,000	12,000	7,600	4
20News	14,609	1,825	1,825	6
UCINews	26,008	2,560	27,556	4
NYT-Topic	19,197	6,400	6,400	9
NYT-Location	19,197	6,400	6,400	10
Yelp	22,800	7,600	7,600	2
Books	20,165	6,719	6,719	8
IMDB	1,600	200	200	2
Twitter	21,983	2,747	2,748	3
MIMIC-III	20,266	2,252	2,252	369

- **BERT** (Devlin et al., 2018) is a fully supervised baseline that trains a transformer model using labeled data.
- **Clinical-BERT** (Alsentzer et al., 2019) is a supervised baseline that trains the BERT model on the clinical text.
- **ConWea**[3] (Mekala and Shang, 2020) expands the keyword vocabulary based on contextual representations of the labels and the corpus.
- **LOTClass**[4] (Meng et al., 2020b) Constructs a keyword vocabulary for pseudo-label generation.
- **X-Class**[5] (Wang et al., 2021) uses clustering to choose the representative documents for each class.
- **ClassKG**[6] (Zhang et al., 2021) iteratively constructs keyword sub-graphs consisting of keywords across data points and derives pseudo-labels by annotating the corresponding sub-graphs.
- **WDDC-MLM**[7] (Zeng et al., 2022) employs a masked language model to generate signal words. They combine the generated words with category names and utilize them for training.
- **NPPrompt**[8] (Zhao et al., 2022) is a zero-shot technique that identifies similar words via non-parametric prompts and uses them as pseudo-labels.
- **MEGClass**[9] (Kargupta et al., 2023) generates pseudo-training labels by iteratively estimating class distribution and contextualized document embeddings.

[3]https://github.com/dheeraj7596/ConWea
[4]https://github.com/yumeng5/LOTClass
[5]https://github.com/ZihanWangKi/XClass
[6]https://github.com/zhanglu-cst/ClassKG
[7]https://github.com/HKUST-KnowComp/WDDC
[8]https://github. com/XuandongZhao/NPPrompt

Evaluation Metrics We use micro-F_1 and macro-F_1 as the evaluation metrics to compare the performance of the text classification methods. More details can be found in Appendix A.

Parameter Settings For each baseline method, we use the default parameter settings as reported in the original papers. More details about the parameter settings of XAI-CLASS can be found in Appendix C.

4.2 Main Results

Our main results are displayed in Table 2. XAI-CLASS outperforms all other baselines on the Yelp, NYT-Topic, Books, and UCINews datasets while providing comparable results on AGNews. We hypothesize our SOTA performance on Yelp is primarily due to its sentimental nature (as it is a polarity dataset) and the label space being distinct (positive or negative sentiment), allowing for there to be more salient words XAI-CLASS can identify compared to other types of datasets used. We provide results on two other polarity datasets, IMDB and Twitter, in Table 3. Our hypothesis is validated by XAI-CLASS outperforming baselines on Yelp and IMDB but not on the Twitter dataset, due to the introduction of the "neutral" class in Twitter.

XAI-CLASS's performance on the Books dataset drastically outperforms all other baselines. We believe this is the result of the indicative and sentiment words that often appear in the description of many books. For example, words commonly found in book descriptions such as "seduce", "murder", and "paranormal" clearly indicate the genres are "romance", "thriller", and "fantasy", respectively.

We believe much of the performance drop-off in 20News is due to labels not being completely disjoint (Zeng et al., 2022). For example, the "electronics" fine-grained class is categorized under the "science" class, although one could argue it would be more appropriate to classify instances of type "electronics" in the "computer" class (Lang, 1995).

[9]https://github.com/pkargupta/MEGClass

Table 2: Micro/macro F_1 scores of baseline methods compared with XAI-CLASS. XAI-CLASS results are based on the optimal number of rounds associated with each dataset. Bolded results correspond to the best-performing model.

Model	Yelp	20News	NYT-Topic	NYT-Loc	Books	AGNews	UCINews
BERT (Supervised)	95.70/95.70	96.60/96.60	95.98/95.01	96.00/95.00	81.00/81.00	93.05/93.06	93.13/93.15
ConWea	71.40/71.20	75.73/73.26	81.67/71.54	85.31/83.81	52.30/52.60	74.43/74.01	32.93/32.69
LOTClass	87.40/87.20	73.78/72.53	67.11/43.38	58.49/58.96	19.90/16.10	86.59/86.56	73.20/72.36
X-Class	86.80/86.80	73.17/73.07	79.01/68.62	89.51/89.68	53.60/54.20	85.74/85.66	68.85/69.62
ClassKG	91.20/91.20	81.00/82.00	72.06/65.76	86.84/83.35	55.00/54.70	**88.80/88.80**	N/A
WDDC-MLM	81.20/81.10	81.21/68.82	81.50/69.20	88.84/86.91	53.86/53.75	88.26/88.25	81.50/81.34
NPPrompt	81.20/81.10	68.90/68.80	64.60/64.20	53.90/53.80	49.60/49.70	85.20/85.20	N/A
MEGClass	87.41/87.41	**81.72/80.63**	85.42/68.03	**93.06/91.93**	56.35/55.71	N/A	N/A
XAI-CLASS	**95.45/95.45**	75.29/71.30	**88.39/80.35**	82.50/86.52	**70.56/70.67**	88.20/88.15	**83.95/83.87**

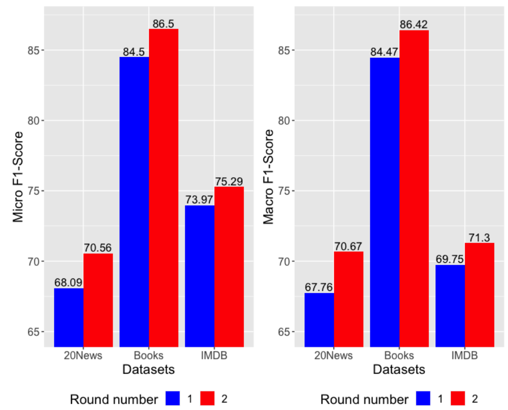


Figure 3: Micro F_1 and macro F_1 scores of two rounds of XAI-CLASS on 20News, Books, and IMDB test sets.

Table 3: F_1 scores of BERT baseline against XAI-CLASS variants. XAI-CLASS-FS is the fully supervised version (with respect to saliency labels) of XAI-CLASS, consisting of ground truth salient labels.

Model	Dev	Test
Dataset: IMDB		
BERT (Supervised)	85.90	85.60
XAI-CLASS-FS	89.50	**87.80**
XAI-CLASS	**91.50**	86.40
Dataset: Twitter		
BERT (Supervised)	77.20	78.10
XAI-CLASS-FS	**78.40**	**79.20**
XAI-CLASS	61.20	63.40

4.3 Ablation Study

Iterative Mutual Enhancement Effectiveness To determine the effectiveness of iterative mutual enhancement, we identify the performance of datasets across multiple rounds. Figure 3 shows these results, clearly indicating that the performance increases when iterating up to a specified number of rounds. It should be noted that the optimal number of rounds is dependent on the dataset, with datasets that have high performance without many rounds most likely requiring fewer rounds than otherwise.

Analyzing Salient Token Utility To analyze the utility of incorporating salient tokens in XAI-CLASS, we conduct experiments on the IMDB and Twitter datasets (Table 3) as they have ground truth salient labels available. Results on both datasets indicate the XAI-CLASS-FS, a variant of XAI-CLASS that includes ground truth saliency labels during training, outperforms XAI-CLASS. This performance increase when utilizing ground truth saliency tokens justifies utilizing salient tokens as it shows that the gold-standard ground-truth saliency labels are being incorporated. The dramatic performance increase when incorporating ground truth salient tokens for Twitter leads us to hypothesize that there's more of a need for proper pseudo-salient representation for datasets that have labels with limited salient words, as the majority of XAI-CLASS misclassifications on the Twitter dataset are on data points whose ground truth is the "neutral" class, which doesn't have many indicative salient words.

Backbone Pre-trained Langauge Models In our experiments, we compared multiple pre-trained language models and chose FLAN-T5 (Chung et al., 2022) as $\mathcal{T}^C$ for the text classification label generation, and BERT (Devlin et al., 2018) as $\mathcal{T}^E$ for the explanation label generation. More information regarding our justification for our choice of $\mathcal{T}^C$ and $\mathcal{T}^E$ can be found in Appendix B.

4.4 Explainability Study

To evaluate the explainability of XAI-CLASS over baseline methods, we qualitatively assess the explainability of Clinical-BERT and XAI-CLASS using six explanation techniques: **Saliency** (Si-

Table 4: Explainability of Clinical-BERT and XAI-CLASS using six explanation techniques on five explanation evaluation metrics (HA, CI, F, RC, DC) on MIMIC-III. Results are in Clinical-BERT/XAI-CLASS format.

Method	F	HA	DC	RC	CI
Random	38.45/**38.56**	0.21/**0.24**	0.02/**0.03**	0.06/0.06	0.13/0.13
ShapSampl	**29.43**/29.28	0.56/**0.61**	0.23/**0.25**	0.21/**0.23**	0.13/**0.14**
LIME	**38.00**/37.89	0.31/**0.33**	0.36/**0.39**	0.61/0.61	0.12/**0.14**
Occlusion	23.00/**25.02**	0.55/**0.56**	0.19/**0.21**	0.34/**0.41**	0.12/**0.14**
Saliency$_{\mu}$	**51.01**/49.23	0.57/**0.59**	**0.34**/0.32	0.26/**0.36**	0.14/**0.19**
Saliency$_{L2}$	44.30/44.30	0.31/**0.37**	0.33/**0.39**	0.24/**0.31**	**0.15**/0.13
InputXGrad$_{\mu}$	20.20/**28.73**	0.53/**0.57**	0.41/**0.42**	0.19/**0.18**	0.15/**0.17**
InputXGrad$_{L2}$	48.72/**49.54**	0.22/**0.24**	0.41/**0.43**	**0.22**/0.21	0.15/**0.16**
GuidedBP$_{\mu}$	**36.66**/35.76	**0.37**/0.34	0.40/**0.43**	0.02/**0.04**	**0.13**/0.12
GuidedBP$_{L2}$	**49.31**/48.38	**0.45**/0.43	0.40/**0.43**	0.19/0.19	**0.14**/0.11

Table 5: Sample of instances with incorrect/ambiguous ground truths in the 20News dataset.

Input Text	Class Prediction	Ground Truth	Salient Word Prediction
72 Chevelle SS for sale. [...] I need money for college. [...] 1972 chevelle super sport rebuilt 402 [...] $ 5995.	Sale	Sports	sale, money, sport
[...] key would appear to be cryptographically useless. [...] The same key is used for both encryption and decryption.	Computer	Science	crypto-graphically, encryp-tion, key
What exactly is an IBM 486 SLC processor? Could someone please tell me if the 486 SLC and 486 SLC2 processors IBM is putting in their Thinkpad 700's.	Computer	Science	IBM, processor, 486
Cultural enquiries more like those who use their backs instead of their minds [...] intolerant of anything outside of their group [..] there is no justification for taking away individuals freedom.	Politics	Sports	cultural, freedom

monyan et al., 2014), **InputXGradient** (Kindermans et al., 2016), **Guided Backpropagation** (Springenberg et al., 2014), **Occlusion** (Zeiler and Fergus, 2014), **Shapley Value Sampling** (Castro et al., 2009), and **LIME** (Ribeiro et al., 2016) over five explanation evaluation metrics (Atanasova et al., 2020) **Agreement with Human Rationales** (HA), **Confidence Indication** (CI), **Faithfulness** (F), **Rationale Consistency** (RC), and **Dataset Consistency** (DC) on the MIMIC-III dataset. Details of the above explanation evaluation metrics can be found in a previous study of explanation techniques in text classification (Atanasova et al., 2020). The results in Table 4 demonstrate that XAI-CLASS improved the model explainability by capturing the saliency information during the training process for all explanation evaluation metrics excluding faithfulness. More results on the explainability case study can be found in Appendix D.

4.5 Case Study

We further explore some cases with incorrect/ambiguous ground truths for multiple reasons, depicted in Table 5. The text in the first row of Table 5 is most likely supposed to be assigned to the "sale" class but is instead labeled with the "sports" class as ground truth, most likely because the word "sport" appears in the text. XAI-CLASS predicted the "sale" class, even though it determined that "sport" was a salient token. This suggests that the model is robust to a small number of words dictating the classification prediction. The second row in Table 5 coincides with the cryptograph example in section 4.2, where one could argue all salient words picked up by the model could be categorized under the term "computer", instead of the ground truth "science". The last two rows of Table 5 appear to be mislabelled, as the third row's text talks exclusively about processors and the fourth example talks only about political issues, yet they are labeled as "science" and "sports", respectively.

5 Conclusion

We propose XAI-CLASS, a novel extremely weakly-supervised text classification method that employs a multi-round question-answering process to generate pseudo-training data and trains a multi-task framework that simultaneously learns both text classification and word saliency prediction. XAI-CLASS has superior performance over baselines for both model performance and explainability. Future work includes extending XAI-CLASS to the multi-label setting.

Limitations

XAI-CLASS, although effective, operates under the assumption of a disjoint label space and is not specifically tailored for fine-grained or multi-label text classification tasks. As a result, it may not perform optimally on datasets like 20News, where there are instances where ground truth labels have some degree of overlap. However, exploring weakly-supervised methods for fine-grained, multi-label text classification is an intriguing direction for future research. Furthermore, it's important to note that XAI-CLASS requires careful consideration when selecting the number of rounds of question answering. It is not designed to scale to a large number of rounds, and typically, no more than three rounds are used. This limitation arises because each round involves two queries for the question answering models: one for generating text classification labels and the other for saliency word generation. This process can be computationally expensive, necessitating a mindful balance between computational resources and desired performance.

Ethics Statement

Given our current methodology, we do not anticipate any significant ethical concerns. We have utilized datasets and models from open-source domains, promoting transparency and accessibility of information. Text classification is a well-established task in natural language processing, widely studied and applied in various domains. However, we acknowledge that our architecture relies on PLMs, which may make decisions based on biases present in the training data. Although our experiments have not revealed any apparent performance issues related to bias, it is important to recognize that this observation may be limited to the datasets we have used. It is crucial to remain vigilant and continue exploring ways to mitigate and address biases that may arise from the use of pre-trained models.

References

Eugene Agichtein and Luis Gravano. 2000. Snowball: Extracting relations from large plain-text collections. In *Proceedings of the Fifth ACM Conference on Digital Libraries*, DL '00, page 85–94, New York, NY, USA. Association for Computing Machinery.

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical bert embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78.

Marco Ancona, Enea Ceolini, A. Cengiz Öztireli, and Markus H. Gross. 2017. A unified view of gradient-based attribution methods for deep neural networks. *CoRR*, abs/1711.06104.

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2020. A diagnostic study of explainability techniques for text classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3256–3274, Online. Association for Computational Linguistics.

Sonia Badene, Kate Thompson, Jean-Pierre Lorré, and Nicholas Asher. 2019. Data programming for learning discourse structure. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 640–645, Florence, Italy. Association for Computational Linguistics.

Javier Castro, Daniel Gómez, and Juan Tejada. 2009. Polynomial calculation of the shapley value based on sampling. *Computers & Operations Research*, 36(5):1726–1730.

Ming Wei Chang, Lev Ratinov, Dan Roth, and Vivek Srikumar. 2008. Importance of semantic representation: Dataless classification. In *AAAI-08/IAAI-08 Proceedings - 23rd AAAI Conference on Artificial Intelligence and the 20th Innovative Applications of Artificial Intelligence Conference*, Proceedings of the National Conference on Artificial Intelligence, pages 830–835. 23rd AAAI Conference on Artificial Intelligence and the 20th Innovative Applications of Artificial Intelligence Conference, AAAI-08/IAAI-08 ; Conference date: 13-07-2008 Through 17-07-2008.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Alex A Freitas. 2014. Comprehensible classification models: a position paper. *ACM SIGKDD explorations newsletter*, 15(1):1–10.

Evgeniy Gabrilovich and Shaul Markovitch. 2007. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *Proceedings of the 20th International Joint Conference on Artifical Intelligence*, IJCAI'07, page 1606–1611, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Fabio Gasparetti. 2016. News Aggregator. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5F61C.

Alistair EW Johnson, David J Stone, Leo A Celi, and Tom J Pollard. 2018. The mimic code repository: enabling reproducibility in critical care research. *Journal of the American Medical Informatics Association*, 25(1):32–39.

Priyanka Kargupta, Tanay Komarlu, Susik Yoon, Xuan Wang, and Jiawei Han. 2023. Megclass: Text classification with extremely weak supervision via mutually-enhancing text granularities. *arXiv preprint arXiv:2304.01969*.

Daniel Khashabi, Sewon Min, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Clark, and Hannaneh Hajishirzi. 2020. Unifiedqa: Crossing format boundaries with a single qa system. *arXiv preprint arXiv:2005.00700*.

Pieter-Jan Kindermans, Kristof Schütt, Klaus-Robert Müller, and Sven Dähne. 2016. Investigating the influence of noise and distractors on the interpretation of neural networks. *arXiv preprint arXiv:1611.07270*.

Ken Lang. 1995. Newsweeder: Learning to filter netnews. In Armand Prieditis and Stuart Russell, editors, *Machine Learning Proceedings 1995*, pages 331–339. Morgan Kaufmann, San Francisco (CA).

Dheeraj Mekala and Jingbo Shang. 2020. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 323–333, Online. Association for Computational Linguistics. [link].

Yu Meng, Jiaxin Huang, Guangyuan Wang, Zihan Wang, Chao Zhang, Yu Zhang, and Jiawei Han. 2020a. Discriminative topic mining via category-name guided text embedding. In *Proceedings of The Web Conference 2020*, pages 2121–2132.

Yu Meng, Jiaming Shen, Chao Zhang, and Jiawei Han. 2018. Weakly-supervised neural text classification. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*. ACM.

Yu Meng, Yunyi Zhang, Jiaxin Huang, Chenyan Xiong, Heng Ji, Chao Zhang, and Jiawei Han. 2020b. Text classification using label names only: A language model self-training approach. *CoRR*, abs/2010.07245.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Alexander Ratner, Christopher De Sa, Sen Wu, Daniel Selsam, and Christopher Ré. 2017. Data programming: Creating large training sets, quickly.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. Model-agnostic interpretability of machine learning. *arXiv preprint arXiv:1606.05386*.

Kai Shu, Subhabrata Mukherjee, Guoqing Zheng, Ahmed Hassan Awadallah, Milad Shokouhi, and Susan Dumais. 2020. Learning with weak supervision for email intent detection. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, page 1051–1060, New York, NY, USA. Association for Computing Machinery.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2014. Deep inside convolutional networks: Visualising image classification models and saliency maps.

Yangqiu Song and Dan Roth. 2014. On dataless hierarchical text classification. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*, AAAI'14, page 1579–1585. AAAI Press.

Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. 2014. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*.

Duyu Tang, Bing Qin, and Ting Liu. 2015. Document modeling with gated recurrent neural network for sentiment classification. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1422–1432, Lisbon, Portugal. Association for Computational Linguistics.

Fangbo Tao, Chao Zhang, Xiusi Chen, Meng Jiang, Tim Hanratty, Lance Kaplan, and Jiawei Han. 2018. Doc2cube: Allocating documents to text cube without labeled data. In *2018 IEEE International Conference on Data Mining (ICDM)*, pages 1260–1265.

Mengting Wan and Julian McAuley. 2018. Item recommendation on monotonic behavior chains. In *Proceedings of the 12th ACM conference on recommender systems*, pages 86–94.

Zihan Wang, Dheeraj Mekala, and Jingbo Shang. 2021. X-class: Text classification with extremely weak supervision. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3043–3053, Online. Association for Computational Linguistics.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489, San Diego, California. Association for Computational Linguistics.

Omar Zaidan, Jason Eisner, and Christine Piatko. 2007. Using "annotator rationales" to improve machine learning for text categorization. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 260–267, Rochester, New York. Association for Computational Linguistics.

Matthew D Zeiler and Rob Fergus. 2014. Visualizing and understanding convolutional networks. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13*, pages 818–833. Springer.

Ziqian Zeng, Weimin Ni, Tianqing Fang, Xiang Li, Xinran Zhao, and Yangqiu Song. 2022. Weakly supervised text classification using supervision signals from a language model. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2295–2305, Seattle, United States. Association for Computational Linguistics.

Lu Zhang, Jiandong Ding, Yi Xu, Yingyao Liu, and Shuigeng Zhou. 2021. Weakly-supervised text classification based on keyword graph. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2803–2813, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28.

Yunyi Zhang, Fang Guo, Jiaming Shen, and Jiawei Han. 2022. Unsupervised key event detection from massive text corpora. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. ACM.

Xuandong Zhao, Siqi Ouyang, Zhiguo Yu, Ming Wu, and Lei Li. 2022. Pre-trained language models can be fully zero-shot learners. *arXiv preprint arXiv:2212.06950*.

A Evaluation Metrics

We report performance based on the micro and macro F_1 scores, which are defined below.

$$F_1 = 2\frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

$$F_1 \text{ macro} = \frac{1}{n}\sum_{i=1}^{n} F_{1,i}$$

$$F_1 \text{ micro} = \frac{2\sum_{i=1}^{n}\text{TP}_i}{2\sum_{i=1}^{n}\text{TP}_i + \sum_{i=1}^{n}\text{FP}_i + \sum_{i=1}^{n}\text{FN}_i}$$

where TP is true positive, FP is false positive, and FN is false negative. We use the sklearn[10] library to obtain these metrics.

B Pseudo-Label and Text Classification Backbone Analysis

We conduct experiments to identify the most appropriate PLMs for $\mathcal{T}^C$ and $\mathcal{T}^E$. To identify the most appropriate $\mathcal{T}^E$, we conduct zero-shot text classification on 7 datasets (Table 6). Flan-T5 performs better than all other models across the seven datasets, indicating why we chose Flan-T5 as the backbone PLM for $\mathcal{T}^C$.

We perform a similar experiment to identify the most appropriate backbone for $\mathcal{T}^E$. Concretely, we perform zero-shot salient label prediction using the IMDB and Twitter datasets, as these are the only datasets we've experimented with that have ground truth saliency labels (Table 7). The results show that BERT and Unified-QA (Khashabi et al., 2020) should be the $\mathcal{T}^E$ backbone of choice when using the IMDB and Twitter datasets for training, respectively.

C Parameter Settings

Runtime Analysis We conduct all of our experiments on an NVIDIA DGX A100 GPU (640GB). The run times for optimal configurations across all datasets can be found in Table 8.

Hyper-parameters The optimal hyper-parameters for our results in Tables 2 and 3 are listed in Table 9. The possible values each of the hyper-parameters can take are listed below:

- $\mathcal{T}^C \in$ {FLAN-T5-SMALL, FLAN-T5-BASE, FLAN-T5-LARGE, FLAN-T5-XL, FLAN-T5-XXL}
  - PLM for psuedo-text classification label generation
- $\mathcal{T}^E \in$ {BERT-BASE, BERT-LARGE, UNIFIED-QA-LARGE, UNIFIED-QA-3B}
  - PLM for psuedo-saliency label generation
- $\lambda \in \{0.5, 0.7, 0.9\}$
  - Hyper-parameter for determining how much of the saliency loss should be incorporated
- Round # $\in \{0, 1, 2, 3\}$
- Learning Rate $\in \{2e-04, 2e-05, 5e-05\}$
- Dropout $\in \{0.1, 0.2, 0.3, 0.4\}$
- Number of Epochs $\in \{1, 2, 3\}$

We implement the PLMs in Python using the HuggingFace Transformer library[11].

D Explanability Case Study

To further evaluate the explainability of XAI-CLASS over the baseline methods, we qualitatively assess the explainability of Clinical-BERT and XAI-CLASS using attention distribution (heatmap). The results in Figure 4 demonstrate that XAI-CLASS improved the model explainability by capturing the saliency information during the training process, particularly in all evaluation metrics excluding faithfullness. The results align well with human-given ICD-9 codes as the explanation for the DRG code prediction.

[10] https://scikit-learn.org/stable/

[11] https://github.com/huggingface/transformers

Table 6: Zero-shot text classification label generation micro/macro F_1-scores across multiple $\mathcal{T}^C$ models. Flan-T5 outperforms all other models across all datasets used, thus serving as our backbone for $\mathcal{T}^C$.

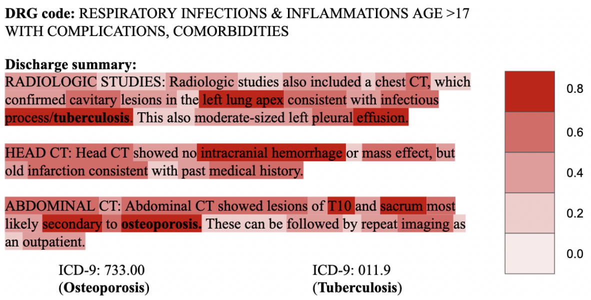
Model	Yelp	20News	NYT-Topic	NYT-Loc	Books	AGNews	UCINews
GPT-2	50.74/42.88	13.97/9.74	7.58/2.72	4.02/2.64	9.88/4.89	26.30/20.16	25.57/11.85
BERT	49.63/39.38	24.99/11.40	31.77/5.58	19.31/4.04	12.56/12.06	26.07/13.18	24.92/11.72
Unified-QA	95.66/95.66	69.75/66.10	76.17/67.30	72.38/75.11	48.13/48.64	86.21/86.12	80.88/80.67
Flan-T5	**97.42/97.42**	**75.34/72.15**	**87.53/78.87**	**81.36/85.90**	**72.03/72.45**	**88.51/88.48**	**84.27/84.16**

Table 7: Zero-shot salient label generation micro/macro F_1-scores across multiple $\mathcal{T}^E$ models on the IMDB and Twitter datasets. We report on these datasets as these are the only datasets with salient labels.

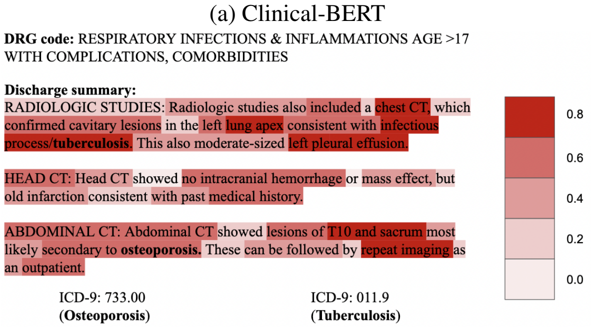
Model	IMDB	Twitter
BERT	**12.26/10.92**	67.56/40.32
Flan-T5	11.26/10.12	69.31/40.94
GPT-2	9.25/8.47	72.16/41.92
Unified-QA	11.57/10.37	**73.8/42.47**

Table 8: Average run time for each dataset for best hyper-parameter configuration.

Dataset	Runtime (hours)
AGNews	10
20News	4
UCINews	4
IMDB	1
Twitter	3



(a) Clinical-BERT



(b) XAI-CLASS

Figure 4: The attention distribution (heatmap) of of Clinical-BERT and XAI-CLASS. A darker red color indicates that the model assigns higher importance to that particular word for explaining the prediction of the DRG code.

Table 9: Optimal hyper-parameters for XAI-CLASS's results in Tables 2 and 3.

Dataset	$\mathcal{T}^C$	$\mathcal{T}^E$	Round #	λ	Learning Rate	Dropout	# Epochs
Books	FLAN-T5-XXL	BERT-BASE	2	0.5	$2e-05$	0.3	3
NYT-Topic	FLAN-T5-XXL	BERT-BASE	1	0.5	$2e-05$	0.3	1
NYT-Location	FLAN-T5-XXL	BERT-BASE	2	0.5	$2e-05$	0.3	3
Yelp	FLAN-T5-XXL	BERT-BASE	1	0.5	$2e-05$	0.3	1
AGNews	FLAN-T5-XXL	BERT-BASE	1	0.5	$2e-05$	0.3	1
20News	FLAN-T5-XL	BERT-BASE	2	0.7	$2e-05$	0.3	3
UCINews	FLAN-T5-XL	BERT-BASE	1	0.5	$2e-05$	0.3	1
IMDB	FLAN-T5-XL	BERT-BASE	1	0.9	$2e-05$	0.4	3
Twitter	FLAN-T5-XL	BERT-BASE	0	0.7	$2e-05$	0.1	3