Verifying effectiveness of `KeyClass` for clinical notes categorization

Project Report

Sayar Ghosh Roy, Atharv Chandratre

{sayar3, atharvc2}@illinois.edu

Group ID: 18

Paper ID: 103

Code: https://github.com/CS-598-DL4H-Spring-2023-Group-18/Verifying-Keyclass

Video: https://youtu.be/KBATJu55oqc

Notebook: https://colab.research.google.com/drive/1mjFFoE7jwR9bNADCzM8pwRl4JPrI9NGX

1 Introduction

The primary aim of our selected paper (Gao et al., 2022), titled 'Classifying Unstructured Clinical Notes via Automatic Weak Supervision' is to alleviate the practice of manual diagnostic coding of clinical notes, a process that is time-consuming, expensive, and error-prone. To address this problem, the paper introduces `KeyClass`, a weakly-supervised text classification framework. `KeyClass` learns a text classification model from class label descriptions alone, eliminating the requirement for manually tagged documents. To assign code labels to specific texts, `KeyClass` leverages pretrained language models and data programming frameworks. During the classification process, `KeyClass` creates certain interpretable heuristics based on keywords extracted from the available text data. Through the adoption of domain-specific language models, the `KeyClass` framework could also be tailored to classification tasks in other highly specialized domains. The comparison between `KeyClass` and other seminal weakly supervised text classification architectures demonstrates `KeyClass`'s efficiency and adaptability.

2 Scope of reproducibility

Our primary goal was to reproduce the experiments described in Gao et al. (2022) involving the `MIMIC-III` clinical notes dataset. Within that, we planned to study the performance achieved by the `KeyClass` model on the `MIMIC-III` clinical notes multi-label classification dataset. To achieve the same, we planned to train the `KeyClass` model on the `MIMIC-III` clinical notes data from scratch (since no pre-trained model or intermediate model checkpoints are provided by the authors). We also planned to validate the category-specific metrics of `KeyClass`.

The paper also presents a study comparing the performance of `KeyClass` against other seminal weakly supervised text classification methods such as `Dataless` (Chang et al., 2008), `WeSTClass` (Meng et al., 2018), and `LOTClass` (Meng et al., 2020) on general domain text classification datasets (such as AG News, DBPedia, IMDB, Amazon user reviews). This was beyond the scope of our evaluation since it does not involve the use of the `MIMIC-III` dataset (or data from a source within the healthcare domain).

2.1 Claims that we evaluated

- The `KeyClass` model effectively mines category-indicative terms from each ICD-9 category
- The `KeyClass` model is capable of achieving an average F1-score greater than the previous state-of-the-art (namely, `FasTag`) for the weakly supervised multi-label classification task on `MIMIC-III` clinical notes data
- (Additional Ablation) The proposed self-training (refining the downstream classifier using the complete training dataset) helps the overall classification performance on the `MIMIC-III` clinical notes dataset

As noted in the final point, we performed an ablation study to test out exactly how much the proposed self-training (refining the downstream classifier using the *complete* training dataset) helps the overall classification performance, specifically for the `MIMIC-III` dataset. The original paper highlights the benefits of self-training for general domain document classification tasks *only*. We were able to *quantify* the performance boost obtained due to self-training for classifying clinical notes.

2.2 Deviations from the original formulation

The authors of the original paper had access to a machine having 187 GB of RAM with multiple

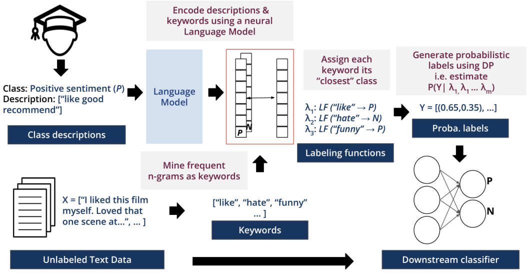
Encode descriptions & keywords using a neural Language Model
Class: Positive sentiment (P)
Description: ["like good recommend"]
Class descriptions
Language Model
P
N
Assign each keyword its "closest" class
λ1: LF ("like" → P)
λ2: LF ("hate" → N)
λ3: LF ("funny" → P)
Labeling functions
Generate probabilistic labels using DP i.e. estimate P(Y| λ1, λ1 … λm)
Y = [(0.65,0.35), …]
Proba. labels
Mine frequent n-grams as keywords
X = ["I liked this film myself. Loved that one scene at…", …]
["like", "hate", "funny" …]
Keywords
P
N
Unlabeled Text Data
Downstream classifier

Figure 1: An overview of the methodology. KeyClass classifies documents from only the class descriptions without any labeled data. (Source: Gao et al. (2022))

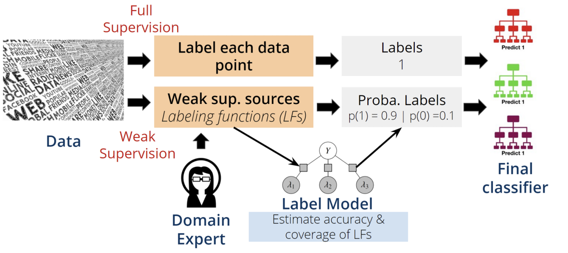

Figure 2: KeyClass removes the need for annotating clinical notes by human experts by bringing in weak supervision from external sources. It creates labeling functions obtained through keyword-matching rules extracted from reference data. (Source: Gao et al. (2022))

total size	333 MB
# train	39541
# test	13181
# categories	19

Table 1: Key dataset statistics

NVIDIA-RTX 2080 GPUs. However, we *only* utilized a single GPU machine with 48 GB of memory. In our proposal and draft, we planned to make some simplifying assumptions for our final experiments. We had initially planned to reduce the number of layers in the fully connected layer and consider a smaller number of classes, say 8, as opposed to 19. We believed that this would allow us to reasonably estimate the efficacy of KeyClass (whether it is capable of accurately performing multi-label classification of clinical notes in a weakly-supervised fashion) within our computational means. We had also noted that certain classes in the MIMIC-III clinical notes dataset are not prevalent enough. For example, 'Pregnancy & childbirth complications' has a prevalence of 0.003. And hence, we planned to focus on the most prevalent classes.

For our final experiments, we modified the classification formulation to a setting that would meet our computational requirements. Instead of using a 19-dimensional multi-hot output vector (which turned out to be infeasible even with a batch size of 1), we kept the overall KeyClass architecture unchanged and swapped out the classification head for 19 separate *one-versus-not* classifiers. This required us to train the model longer (19 times for the 19 classification heads). But it fit into our memory budget. Also, this allowed us to train the model for *all* of the 19 classes as opposed to a smaller subset of classes. As a consequence, for fine-tuning the downstream classifier, we had to utilize the available training labels as opposed to the generated pseudo-labels (since pseudo-labels for the *not-in-class* case would not be well-defined), thereby changing the formulation from weakly-supervised to semi-supervised.

With these changes, we were able to verify both the category-indicative keyword mining aspect of KeyClass as well as the categorization efficacy of the overall KeyClass framework considering all of the 19 categories.

3 Methodology

Firstly, we communicated with the authors of the original paper to confirm that they did not share the code for the multi-label version of KeyClass and that we would need to implement the same. The authors advised us to thoroughly study the shared code for the multi-*class* KeyClass model capable of performing weakly-supervised classification on general-domain text classification datasets, and then implement the multi-label version based on the same. Therefore, we re-implemented the code for model definition (mainly in models.py) and training (mainly in train_models.py) specifically for the multi-label categorization setting. We were unable to run the multi-label version of the model for the 19 classes using the complete MIMIC-III dataset. Hence, we additionally implemented the proposed 19-way *one-versus-not* classifier setting.

In the following subsection, we shall discuss the overall flow of the KeyClass model.

3.1 Model description

The aim of KeyClass is to categorize documents *without* any labeled data, leveraging class descriptions only (Figure. 1). To achieve this, it automatically creates labeling functions (LFs) which, along with data programming, generate probabilistic labels for unlabeled training data (Figure. 2). These weak labels are then used to train a downstream classifier. Broadly, the KeyClass archi-

Category Name	#Train	#Test
Infections and Parasitic	10664	3548
Neoplasms	6458	2121
Endocrine, Nutritional, and Metabolic	25947	8653
Blood and Blood Forming Organs	14286	4720
Mental Disorders	11851	3880
Nervous System	10031	3354
Sense Organs	2790	918
Circulatory System	30945	10390
Respiratory System	18423	6154
Digestive System	15348	5082
Genitourinary System	15950	5303
Pregnancy and Childbirth Complications	113	43
Skin and Subcutaneous Tissue	4453	1471
Musculoskeletal System and Connective Tissue	7367	2441
Congenital Anomalies	2070	735
Perinatal Period Conditions	2819	894
Injury and Poisoning	14598	4900
External Causes of Injury	16436	5429
Supplementary	27319	9107

Table 2: # train and test data points in each category

tecture could be subdivided into three main steps. First, the above-discussed label model generates probabilistic labels for individual documents. Second, a downstream classifier is *pre*-trained using those weak labels. More specifically, the downstream classifier is trained using only the top-K most confident document to label mappings generated in step 1. Finally, the third step is that of *self*-training. Here, the complete architecture (encoder plus downstream classifier combination) is trained on the entire training dataset to further refine the model parameters.

We now discuss specific modules within the KeyClass architecture. For mathematical formulation of the problem setting, we refer the reader to Gao et al. (2022) due to lack of space.

Data Programming for Weak Text Classification KeyClass removes the need for annotating large quantities of clinical notes data by human experts by bringing in weak supervision from pre-trained Large Language Models (LLMs). This is done using labeling functions that are in turn obtained through keyword-matching rules that are automatically extracted from reference data.

Finding Class Descriptions In order to initiate the unsupervised classification procedure, KeyClass leverages natural language descriptions for each data class in a dataset. For MIMIC-III clinical notes classification, we obtain the descriptions of the various ICD-9 classes provided by domain experts. In our implementation, we mined descriptions for each of the 19 ICD-9 categories from Wikipedia (as advised by the authors of Gao et al. (2022)).

Finding Relevant Keywords The KeyClass architecture automatically discovers highly relevant keyphrases for each class. The module first obtains frequent n-grams from the training corpus to serve as candidate keyphrases. Then, it maps each candidate keyphrase to the most semantically related category description using a pre-trained Language Model. A keyphrase is finally assigned to its semantically closest category based on cosine similarity.

Label Model Parameters	Value
num epochs	100
learn rate	$1e-2$

Downstream Model Parameters	Value
num epochs	20
learn rate	$1e-4$
patience	3
top K	50

Self-training Parameters	Value
learn rate	$1e-6$
patience	8

Table 3: Key hyperparameters and their default values

Probabilistically Label Data KeyClass constructs a labeling function vote matrix and generates probabilistic labels for all training documents using the label model.

Fine-tune Downstream Text Classifier Lastly, KeyClass fine-tunes a downstream classifier using the rich feature representations provided by the neural Language Model. Initially, KeyClass trains the downstream classifier using the top k documents with the most confident label estimates. The model is then *self*-trained on the complete dataset as the final refinement step.

3.2 Data Description

The dataset in question is that of MIMIC-III clinical notes classification. We registered ourselves on the PhysioNet[1] and received access to the raw MIMIC-III data. We then extracted, preprocessed, and built the clinical notes text categorization dataset.

We first isolated the DIAGNOSES_ICD.csv and NOTEEVENTS.csv files within MIMIC-III. We then created a notes-to-ICD_codes mapping by right-joining the data in these two csv files using the HADM_ID (following the procedure outlined in Venkataraman et al. (2020)). Additionally, we cleaned the textual data (utilizing the cleantext library in Python). We then used the same 70:30 train-test split used in Venkataraman et al. (2020). We finally encoded our target variable as n-dimensional multi-hot vectors (1 corresponding to every diagnosis code for a patient). The key dataset statistics can be found in Table 1. We have additionally aggregated the class-specific data distribution in Table 2.

[1] https://physionet.org/

Category Terms	Mined Keywords
infections, parasitic	also infect, appear infect, bacteri infect, bacteria patient, chronic infect, concern infect, concern infecti, concern wound infect, develop infect, discuss infecti, due infecti, experi fever, experi follow fever, fever infect, fever infecti, followup infecti, found infect, fungal infect, fungal rash, infect appear, infect chronic, infect clinic, infect concern, infect due pneumophila, infect fever, infect patient, infect present, infect skin, infect wound, infect wound site, infecti, infectionsepsi, like infect, like infecti, may infecti, medic infect, pathogen, patient infect, periorbit ecchymosi, seen infecti, skin infect, suggest infecti, tender drainag fever, tract infect, tract infect also, tract infect pleas, treat antibiot infect, viral infect, well infecti, wound infect
neoplasms	brain metastas, brain metastasi, cancer metastas, cancer metastat, cancer recent, carcinoma, carcinoma metastat, cell carcinoma, cell carcinoma left, cell carcinoma right, cell carcinoma sp, cell tumor, consist adenocarcinoma, extens metastat, extens tumor, hematoma like, known metastat, like metastat, lymphangit carcinomatosi, malign cell, malign cell consist, metamyelocyt, metaplasia, metastas, metastas patient, metastat carcinoma, metastat lesion, metastat melanoma, metastat nonsmal cell, mycoplasma, neg metastat, neobladd, neomycin, neoplasia, neoplasm, neoplast, neoplast process, neosur, neuroma, nonsmal cell carcinoma, patient noncompli, rectu sheath hematoma, show malign cell, show tumor, suggest metastat, tumor also, tumor cell, tumor major, tumor seen, wide metastat
endocrine, nutritional, metabolic	blood sugar admiss, blood sugar also, blood sugar meal, cardiac diabet diet, cardiac diet, cardiac healthi diet, diet, diet abl, diet also, diet cardiac, diet consist, diet discharg, diet discharg instruct, diet electrolyt, diet exercis, diet follow, diet followup, diet followup instruct, diet monitor, diet per, diet physic, diet postop, diet seen, diet throughout, diet use, diet well, electrolyt nutrit birth, endocrinologist, follow diet, follow endocrinologist, food diet, healthi diet, heart healthi diet, hyperglycemia, hyperglycemia discharg, hyperglycemia like, hyperglycemia major, hypertriglyceridemia, hyperuricemia, lipid, normal healthi diet, note blood sugar, present hyperglycemia, recommend diet, regular diet adequ, regular diet discharg, regular diet postop, sodium diet followup, steroid induc hyperglycemia, well diet
blood, blood-forming organs	abneg blood, also blood, blood, blood also, blood asaneg, blood blood, blood cell, blood cell also, blood cell blood, blood cell differenti, blood cell due, blood cell given, blood cell hematocrit, blood cell hospit, blood cell occasion, blood cell postop, blood cell transfus, blood cell unit, blood discharg, blood layer, blood like, blood product, blood product within, blood rectal, blood rectum, blood throughout, blood transfus blood, blood vessel, blood vessel turn, blood within, bloodfung, bloodstream, bloodting, cell blood, complet blood, consist blood, develop blood, episod blood, final growth blood, function blood, growth blood, help blood, hematocrit blood, inpati blood, intact blood, like blood, process blood, system blood, th blood, whole blood

Table 4: Mined keyphrases for each category. We show only the first 4 categories here due to lack of space. For a complete list of all mined keyphrases, refer to the Appendix (Section 6).

3.3 Hyperparameters

We have listed the main hyperparameters for each module along with their default values in Table 3.

3.4 Implementation

We used `paraphrase-mpnet-base-v2` as the base text encoder. We utilized a 2-layer multi-layer perceptron (MLP) with hidden layer sizes: 768 and k, where k is the number of categories. For our implementation of the multi-label classification head, we used `LeakyReLU` activation with binary cross-entropy with logits loss as the criterion, as suggested in the original paper. For our implementation of the 19-way *one-versus-not* classifier, we utilized the same MLP structure but with `sigmoid` activation (as it consistently gave lower training losses during our initial runs) and used the cross-entropy loss during training.

For finding class descriptions, we mined descriptions for each of the 19 broad ICD-9 categories from `Wikipedia`, following Gao et al. (2022). We utilized the `cleantext` Python library for pre-processing the clinical notes data following Venkataraman et al. (2020).

Our implementations of `KeyClass`, with both the multi-label and the k-way *one-versus-not* classification heads, are generic. In that, they can perform multi-label classification on any number of classes (19 for the `MIMIC-III` clinical notes data).

For additional insights into some of our implementation choices, refer to Subsection 5.1 (Communication with original authors). Our codebase is publicly available here[2].

3.5 Computational requirements

According to the authors of Gao et al. (2022), the models were trained on a machine having 40 Intel Xeon Silver 4210 CPUs and 4 NVIDIA RTX2080 GPUs with a total memory of 187 GB. As we did not have access to such a powerful server, we made some simplifying assumptions for our final experiments (described in the Section 2). This is in line with the initial estimations we made in our proposal and draft.

We initially tested out our implementations on a `Google Colab` GPU instance (having roughly 13 GB of usable RAM) using a dataset subsample with 4 categories. As noted before, we only had access to one machine with 48 GB of memory, which would not be enough to train the multi-label `KeyClass` model with 19-categories (specifically the fine-tuning the downstream classifier step). Note that we did attempt the same with a batch size of 1 and that led to a `CUDA out of memory` error.

Thus, we implemented a modified classification formulation for our final *training the downstream classifier* experiments (details in Section 2.2). This allowed us to train the architecture end-to-end on the entire dataset considering all of the 19 classes (as opposed to a subset of them). We could successfully complete mining category specific keywords

[2] `https://github.com/CS-598-DL4H-Spring-2023-Group-18/Verifying-Keyclass`

ID	Class Name	F1	Precision	Recall
0	infectious and parasitic diseases	0.134 ± 0.003	0.087 ± 0.002	0.294 ± 0.004
1	neoplasms	0.215 ± 0.004	0.149 ± 0.003	0.386 ± 0.004
2	endocrine, nutritional, metabolic	0.103 ± 0.003	0.065 ± 0.002	0.254 ± 0.004
3	blood, blood-forming organs	0.565 ± 0.005	0.477 ± 0.006	0.691 ± 0.004
4	mental disorders	0.600 ± 0.025	0.516 ± 0.027	0.718 ± 0.019
5	nervous system	0.514 ± 0.021	0.424 ± 0.022	0.651 ± 0.017
6	sense organs	0.182 ± 0.020	0.123 ± 0.016	0.350 ± 0.022
7	circulatory system	0.696 ± 0.005	0.622 ± 0.006	0.789 ± 0.004
8	respiratory system	0.723 ± 0.023	0.655 ± 0.027	0.809 ± 0.016
9	digestive system	0.009 ± 0.001	0.005 ± 0.000	0.068 ± 0.002
10	genitourinary system	0.240 ± 0.004	0.170 ± 0.004	0.412 ± 0.004
11	pregnancy, childbirth complications	0.006 ± 0.000	0.003 ± 0.000	0.056 ± 0.002
12	skin, subcutaneous tissue	0.297 ± 0.005	0.218 ± 0.004	0.467 ± 0.005
13	musculoskeletal system, connective tissue	0.202 ± 0.004	0.138 ± 0.003	0.372 ± 0.005
14	congenital anomalies	0.009 ± 0.001	0.005 ± 0.000	0.069 ± 0.002
15	perinatal period conditions	0.000 ± 0.000	0.000 ± 0.000	0.003 ± 0.001
16	injury and poisoning	0.058 ± 0.002	0.035 ± 0.001	0.186 ± 0.003
17	external causes of injury	0.022 ± 0.001	0.012 ± 0.001	0.111 ± 0.003
18	supplementary	0.231 ± 0.004	0.162 ± 0.003	0.402 ± 0.004

Table 5: Ablation Study: Metrics achieved by the `KeyClass` model *without* self-training

(plus creating the label model & vote matrix). And we were able to obtain results both for initial pre-training and fine-tuning the downstream classifier. We had to update our problem setting from that of weakly-supervised to semi-supervised (as we had to use some true training labels for fine-tuning the downstream classifier).

For number of training epochs, refer to Table 3. The total time for doing a complete run (pre-training plus fine-tuning the downstream classifier for all of the 19 classes) was around 41 hours with a batch size of 16. The average time taken for each epoch for training the downstream classifier was around 38 minutes.

4 Results and Analysis

In this section, we analyze the outcomes of our conducted experiments.

4.1 Mining Category indicative keyphrases

We share our outputs of the category-indicative keyphrase generation step of `KeyClass`. We trained the label model and created the labeling function vote matrix with the 19 categories considering the complete training dataset. In Table 4, we have included the category specific keyphrases for each class (only 4 within the main report due to lack of space) generated from the label model and vote matrix creation step. Overall, we see that the model manages to find certain novel keyphrases while exhaustively aggregating all variations of a particular keyphrase seen in the text corpora.

4.2 Pre self-training phase (Ablation Study)

In Table 5, we report the metrics achieved *prior to* `KeyClass` self-training. Here, we observe that the category 'respiratory system' is the simplest to classify (F1 score of 0.723) for the self-trained model while the class 'perinatal period conditions' is the toughest (F1 score of 0). For this setting, we have average F1, precision, and recall of 0.253, 0.203, and 0.373 respectively. We observe that utilizing labeling functions and *pre*-training using confident label assignments alone is not sufficient to reach a respectable average F1 score. Thus, to classify clinical notes, it is essential to have some form of fine-tuning of the complete architecture.

4.3 Fine-tuning Downstream Classifier

In Table 6, we report the observed metrics after fine-tuning the downstream classifier. Here, we observe that the category 'genitourinary system' is the simplest to classify (F1 score of 0.995) while the class 'perinatal period conditions' is the toughest (F1 score of 0.075). The mean F1, precision, and recall observed after fine-tuning were 0.576, 0.675, and 0.683 respectively.

The original paper reported an average F1 score of 0.625 for weakly supervised `KeyClass` and 0.525 for weakly supervised `FasTag`. While we obtain an average F1 score of 0.576 with our variation of `KeyClass`. This drop in performance (from the original `KeyClass` model) is expected because of our compulsion to utilize 19 separate classifier heads as opposed to a 19-dimension multi-hot category-indicator vector. In the original `KeyClass` setting, the joint training for all categories forces the model to learn the cross-category interactions. However, in our simplification, each category is treated independently and the MLP fails to capture the rich inter-category relationship signals. The original paper reported an average precision of 0.507 and an average recall of 0.896. Thus, we notice that we obtained a higher precision but a lower recall compared to the values reported in the

ID	Class Name	F1	Precision	Recall
0	infectious and parasitic diseases	0.899 ± 0.003	0.869 ± 0.004	0.932 ± 0.002
1	neoplasms	0.917 ± 0.003	0.892 ± 0.004	0.944 ± 0.002
2	endocrine, nutritional, metabolic	0.584 ± 0.005	0.499 ± 0.005	0.706 ± 0.004
3	blood, blood-forming organs	0.510 ± 0.024	0.420 ± 0.024	0.648 ± 0.019
4	mental disorders	0.731 ± 0.004	0.720 ± 0.059	0.815 ± 0.003
5	nervous system	0.597 ± 0.025	0.512 ± 0.027	0.715 ± 0.019
6	sense organs	0.897 ± 0.003	0.866 ± 0.004	0.930 ± 0.002
7	circulatory system	0.448 ± 0.005	0.756 ± 0.041	0.598 ± 0.004
8	respiratory system	0.468 ± 0.005	0.759 ± 0.039	0.615 ± 0.004
9	digestive system	0.638 ± 0.005	0.557 ± 0.006	0.746 ± 0.004
10	genitourinary system	0.995 ± 0.001	0.993 ± 0.001	0.997 ± 0.001
11	pregnancy, childbirth complications	0.728 ± 0.023	0.660 ± 0.027	0.812 ± 0.017
12	skin, subcutaneous tissue	0.183 ± 0.019	0.124 ± 0.015	0.351 ± 0.021
13	musculoskeletal system, connective tissue	0.371 ± 0.005	0.577 ± 0.145	0.533 ± 0.004
14	congenital anomalies	0.836 ± 0.004	0.790 ± 0.005	0.889 ± 0.003
15	perinatal period conditions	0.075 ± 0.002	0.825 ± 0.079	0.212 ± 0.003
16	injury and poisoning	0.147 ± 0.004	0.779 ± 0.069	0.310 ± 0.005
17	external causes of injury	0.437 ± 0.005	0.604 ± 0.128	0.589 ± 0.004
18	supplementary	0.486 ± 0.005	0.628 ± 0.115	0.629 ± 0.004

Table 6: Metrics achieved by the `KeyClass` model after fine-tuning the downstream classifier

original paper.

These metrics showcase the efficacy of the `KeyClass` architecture (mining category indicative keywords, data programming, creating labeling functions, training the downstream classifier). Our simplified version beats the second best weakly-supervised setting for the task, namely `FasTag` — even with our restrictions, we obtained a 5.10% jump over weakly-supervised `FasTag` proving the efficacy of `KeyClass`' modeling decisions.

5 Discussion

In this section, we provide an overview of our results and their implications. Firstly, as stated in Section 4, we showcased the efficacy of the overall workflow of `KeyClass` that includes steps such as mining category indicative keywords, data programming, creating labeling functions, and training the downstream classifier. Our simplified version with a modified classifier head achieved a 5.10% jump over `FasTag` proving the validity of the design choices made in `KeyClass`. The F1 score that we observed expectedly fell short of the numbers reported in the original paper as we had to make some simplifications to the formulation (due to our computational constraints). However, we were able to train our implementation end-to-end, mine category specific keywords, and obtain results for the complete 19-class multi-label classification. Overall, the data processing was straightforward. We had to verify some of the intermediate outputs and build a robust data processing pipeline for the raw `MIMIC-III` dataset. Modifying the `KeyClass` code for the multi-label categorization setting was time-consuming as we had to deal with a series of cascading errors. However, the biggest challenge was that of coming up with a compromise on the overall formulation such that we could conduct all experiments considering all of the 19 classes (end-to-end training using the complete dataset). With our 19-way *one-versus-not* classification scheme, the run-time of experiments became much longer but we were able to fit the model into memory (the most expensive step being that of fine-tuning the complete architecture).

The original paper (Gao et al., 2022) was very well written and provided all of the key details. The only downsides were that the processed dataset and the implementation for the multi-label version of `KeyClass` were not publicly available. However, we understand that sharing such data could be tricky since the source data (`MIMIC-III`) is not openly available). The provided codebase was well documented and structured sensibly.

5.1 Communication with original authors

As stated earlier, we communicated with the primary authors of the original paper, namely, Chufan Gao and Mononito Goswami. They confirmed that they did *not* share the code for the multi-label classification scheme of `KeyClass`. They advised us to thoroughly study the existing codebase and implement the multi-*label* version of `KeyClass` based on the same. They further confirmed that they currently do not possess the pre-processed `MIMIC-III` clinical notes dataset or any pre-trained models for `MIMIC-III` clinical notes classification (they only have access to trained models for the generic text classification datasets). Chufan shared some tips and pointers for (a) thresholding the multi-label classifier, (b) creating the seed category descriptions for the label model, (c) building the data processing pipeline, and (d) designing the full-scale experiments. We would like to thank Chufan and Mononito for their guidance.

References

Ming-Wei Chang, Lev-Arie Ratinov, Dan Roth, and Vivek Srikumar. 2008. Importance of semantic representation: Dataless classification. In *AAAI Conference on Artificial Intelligence*.

Chufan Gao, Mononito Goswami, Jieshi Chen, and Artur Dubrawski. 2022. Classifying unstructured clinical notes via automatic weak supervision.

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*, pages 983–992. ACM.

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

Guhan Ram Venkataraman, Arturo Lopez Pineda, Oliver J. Bear Don't Walk Iv, Ashley M. Zehnder, Sandeep Ayyar, Rodney L. Page, Carlos D. Bustamante, and Manuel A. Rivas. 2020. Fastag: Automatic text classification of unstructured medical narratives. *PloS One*, 15(6):e0234647.

6 Appendix

In this section, we include additional details that we had to omit in the main draft due to lack of space. Currently, we include a subsection listing the category-indicative keyphrases mined from the KeyClass model.

6.1 Mined keyphrases

In this subsection, we list out the mined category-indicative keyphrases for each of the 19 classes in our dataset.

Category Terms	**Mined Keywords**
infections, parasitic	also infect, appear infect, bacteri infect, bacteria patient, chronic infect, concern infect, concern infecti, concern wound infect, develop infect, discuss infecti, due infecti, experi fever, experi follow fever, fever infect, fever infecti, followup infecti, found infect, fungal infect, fungal rash, infect appear, infect chronic, infect clinic, infect concern, infect due pneumophila, infect fever, infect patient, infect present, infect skin, infect wound, infect wound site, infecti, infectionsepsi, like infect, like infecti, may infecti, medic infect, pathogen, patient infect, periorbit ecchymosi, seen infecti, skin infect, suggest infecti, tender drainag fever, tract infect, tract infect also, tract infect pleas, treat antibiot infect, viral infect, well infecti, wound infect
neoplasms	brain metastas, brain metastasi, cancer metastas, cancer metastat, cancer recent, carcinoma, carcinoma metastat, cell carcinoma, cell carcinoma left, cell carcinoma right, cell carcinoma sp, cell tumor, consist adenocarcinoma, extens metastat, extens tumor, hematoma like, known metastat, like metastat, lymphangit carcinomatosi, malign cell, malign cell consist, metamyelocyt, metaplasia, metastas, metastas patient, metastat carcinoma, metastat lesion, metastat melanoma, metastat nonsmal cell, mycoplasma, neg metastat, neobladd, neomycin, neoplasia, neoplasm, neoplast, neoplast process, neosur, neuroma, nonsmal cell carcinoma, patient noncompli, rectu sheath hematoma, show malign cell, show tumor, suggest metastat, tumor also, tumor cell, tumor major, tumor seen, wide metastat
endocrine, nutritional, metabolic	blood sugar admiss, blood sugar also, blood sugar meal, cardiac diabet diet, cardiac diet, cardiac healthi diet, diet, diet abl, diet also, diet cardiac, diet consist, diet discharg, diet discharg instruct, diet electrolyt, diet exercis, diet follow, diet followup, diet followup instruct, diet monitor, diet per, diet physic, diet postop, diet seen, diet throughout, diet use, diet well, electrolyt nutrit birth, endocrinologist, follow diet, follow endocrinologist, food diet, healthi diet, heart healthi diet, hyperglycemia, hyperglycemia discharg, hyperglycemia like, hyperglycemia major, hypertriglyceridemia, hyperuricemia, lipid, normal healthi diet, note blood sugar, present hyperglycemia, recommend diet, regular diet adequ, regular diet discharg, regular diet postop, sodium diet followup, steroid induc hyperglycemia, well diet
blood, blood-forming organs	abneg blood, also blood, blood, blood also, blood asaneg, blood blood, blood cell, blood cell also, blood cell blood, blood cell differenti, blood cell due, blood cell given, blood cell hematocrit, blood cell hospit, blood cell occasion, blood cell postop, blood cell transfus, blood cell unit, blood discharg, blood layer, blood like, blood product, blood product within, blood rectal, blood rectum, blood throughout, blood transfus blood, blood vessel, blood vessel turn, blood within, bloodfung, bloodstream, bloodting, cell blood, complet blood, consist blood, develop blood, episod blood, final growth blood, function blood, growth blood, help blood, hematocrit blood, inpati blood, intact blood, like blood, process blood, system blood, th blood, whole blood

mental disorders	abus bipolar, admiss mental, bipolar disord, bipolar disord present, complaint mental, concern mental, depress bipolar, depress chronic, depress mental, depress patient, depress psychiatri, depress symptom, depressionanxieti, depressionanxieti patient, deterior mental, deterior mental statu, diagnosi alter mental, disord chronic, disord patient, dm neuropathi, due alter mental, histori bipolar disord, ho bipolar, mental, mental health, mental ill, mental retard, mental state, mental statu concern, mental statu worsen, mild mental, mild mental retard, pain mental, paranoid schizophrenia, patient alter mental, patient mental, patient psychiatr, poor mental, psych med, psychiatr ill, psychiatr patient, psychiatri patient, psychotrop, schizophrenia, schizophrenia discharg, schizophrenia present, state symptom, symptom patient, worsen mental, worsen mental statu
nervous system	admiss neurolog, arm neuro, brain also, brain complet, brain obtain, brain perform, brain stem, brain stem reflex, brainstem, brainstem reflex, brainstem respons, enhanc brain, intact motor function, intact neuro, intact neurolog, known brain, left neural, mid brain, motor neuron, neuro cn, neuro cn grossli, neuro cn intact, neuro intact, neuro mental, neuro motor, neuro orient, neuro respond, neuro unrespons, neurocheck, neurolept, neurolog activ, neurolog function, neurolog histori, neurolog respond, neuromuscular, neurophysiolog, normal neuro, per neuro, per neurolog, present neuro, present neurolog, pt neuro cn, right neural, system neuro, system neuro patient, system neurolog, throughout neuro, throughout neurolog, upper motor neuron, within brain
sense organs	appropri sensori hear, bowel sound organomegali, brain find, cryptogen organ, detect, detect organ, detect organ rare, ear, ear canal, ear eye nose, ear nose, ear nose eye, ear nose mouth, enhanc detect organ, find liver, hear finger, hear intact finger, heard abdomen, intact ear, liver lobe, liver seen, lobe like, lobe liver, lobe seen, may detect, normal sensori hear, organ, organ consist, organ gram, organ identifi, organ may, organ system, pass ear psychosoci, respons pass ear, seen brain, seen kidney, seen liver, sensori audiolog hear, sensori hear, sensoryaudiolog, sensoryaudiolog hear, sensoryaudiolog hear screen, solid organ, sound abdomen, sound organomegali, sound posit abdomen, sound present organomegali, sound rectal, viii hear finger, vision hear
circulatory system	biventricular, biventricular function, biventricular systol function, call ventricular, cardiovascular system, circul, circulatori, circulatori arrest, conclus left ventricular, cours system cardiovascular, deep hypotherm circulatori, effect ventricl, flow heart, fluid around heart, good biventricular function, hypotherm circulatori, hypotherm circulatori arrest, interventricular, intraventricular, intraventricular blood, intraventricular conduct, intraventricular conduct delay, moder ventricular, subarachnoid intraventricular, sustain ventricular, system cardiovascular, system vascular, th ventricl, tube cardiovascular, vascular flow, ventcontrol blood, ventricl, ventricl also, ventricl cistern, ventricl consist, ventricl like, ventricl moder, ventricl rv, ventricl ventricular, ventricl well, ventricl without, ventricular, ventricular caviti moder, ventricular chamber, ventricular conduct, ventricular drain, ventricular function, ventricular system, ventriculoperiton, vessel heart
respiratory system	air lung, breath respiratori, breath ventil, chest decreas breath, chest lung, copd oxygen, follow lung, hypotens respiratori, hypox respiratori, like lung, lower lung, lung, lung also, lung area, lung breath, lung brief, lung bronchial breath, lung chest, lung coars breath, lung comfort, lung consist, lung decreas breath, lung expiratori, lung follow, lung function, lung good air, lung like, lung lung, lung poor air, lung process, lung suggest, lung well, oxygen chest, rate lung, respiratori breath, respiratori function, respiratori lung, respiratori tract, system respiratori, throughout lung, throughout right lung, upper lung, upper respiratori, upper respiratori tract, ventil breath, ventil respiratori, ventilatorassoci, ventilatori, way lung, within lung
digestive system	abdomen bowel, abdomen normal bowel, abdomen posit bowel, bowel, bowel also, bowel colon, bowel consist, bowel function, bowel habit, bowel like, bowel movement, bowel regimen, bowel suggest, bowel well, colon small bowel, concern bowel, control bowel, decreas bowel, diet bowel, diet use stool, discharg gram gastrointestin, gastric, gastroc, gastroenter, gastroesophag, gastrointestin abdomen, gastrointestin tract, gastrostomi, gi bowel, good bowel, gram gastrointestin, intestin, mass bowel, need bowel, normal bowel, obes bowel, posit bowel, regular diet bowel, small intestin, stomach, stomach acid, stomach also, stomach content, stomach like, stomach small bowel, stool bowel, stool colon, stool soften eat, whole stomach, within stomach
genitourinary system	andor genit, anu patent genitourinari, bowel sound genitourinari, codomin system, colac gentl, coloni type consist, complic gestat, continu glargin, diures gentli, diuret gentli, facil gentiva, gen, gen somnol, generalnad, genit, genitourinari, genitourinari examin, genitourinari normal, genitourinari normal femal, genitourinari normal male, genl, genta, gentamicin rule, gentamicin rule sepsi, gentiva, gentl, gentl diuresi, gentl fluid, gentl hydrat, gentl iv, gentli, gentli diures, gentli hydrat, gentli pat, genu, given gentl, incis gentli, mass genitourinari, order continu, order continu ambul, part duodenum brief, patent genitourinari, servic facil gentiva, sound genitourinari, sound genitourinari normal, start gentli diures, two coloni morpholog, type coloni, vital gen, vs gen
pregnancy, childbirth complications	complic pregnanc, complic preterm labor, deliveri pregnanc complic, deliveri preterm labor, diagnos prematur infant, diagnos preterm, discharg diagnos preterm, distress birth, due preterm labor, first pregnanc, hospit preterm labor, infant birth, infant hemodynam, labor infant, mother pregnanc, mother pregnanc complic, onset labor, onset preterm labor, postpartum, preeclampsia, pregnanc, pregnanc also, pregnanc born, pregnanc complic preterm, pregnanc conceiv, pregnanc mother, pregnanc notabl, pregnanc reportedli, pregnanc uncompl mother, pregnancyinduc, pregnancyinduc hypertens, pregnant, prenat care, prenat diagnosi, prenat laboratori, present preterm labor, preterm deliveri, preterm infant, preterm infant chronic, preterm labor, preterm labor infant, preterm labor mother, preterm labor treat, preterm low birth, twin pregnanc, uncompl pregnanc, unknown pregnanc, unstopp preterm labor, woman prenat, woman whose pregnanc
skin, subcutaneous tissue	area skin, area skin breakdown, cell skin, extrem skin, extrem skin jaundic, extrem skin rash, intact skin, intraparenchym subarachnoid, lesion skin, lower extrem skin, normal skin, palpat skin subcutan, patient skin, side subdur hematoma, skin cancer, skin cm, skin erythemat, skin fold, skin graft, skin graft left, skin graft right, skin lesion, skin neuro, skin patient, skin rash skin, skin skin, skin subcutan, skin subcutan tissu, skin without lesion, skin wound, split thick skin, subacut subdur hematoma, subarachnoid blood, subcutan emphysema, subcutan emphysema right, subcutan fat, subcutan hematoma, subdur hematoma, subdur hematoma along, subdur hematoma layer, subdur hematoma overli, subdur hematoma subarachnoid, subdur hematoma without, subdur subarachnoid, subgal hematoma, tender skin, thick skin graft, throughout skin, traumat subarachnoid, type consist skin
musculoskeletal system, connective tissue	calf thigh, collagen, collagenas, deep tendon, deep tendon reflex, fibroid social, flexionextens knee, flexionextens knee flexionextens, hip flexionextens knee, hypertens osteoarthr, inner thigh, intercost muscl flap, interspin ligament, knee flexionextens, knee neuro, leg neuro, medial thigh, mucosa, mucosa neck, muscl cramp, muscl pain, muscl strength throughout, muscl use abdomen, musculatur, musculatur symmetr, musculatur symmetr viii, musculoskelet, musculoskelet hip, musculoskelet muscl, musculoskelet muscl wast, musculoskelet normal, musculoskelet normal spine, musculoskelet pain, musculoskelet patient, normal musculoskelet, normal musculoskelet normal, osteoarthr, osteoarthr chronic, osteoarthr sp, osteoarthrit, sacroiliac joint, sever osteoarthr, tender calf thigh, tendon, tendon reflex throughout, thigh neuro, vasculatur, vasculatur within, vertebrobasilar, vertebrobasilar system
congenital anomalies	anemia also, anemia present, anemia sp, anemia unclear, aneurysm also, aneurysm aris, aneurysm follow, aneurysm mm, aneurysm physic, aneurysm present, aneurysm seen, aneurysm sp, aneurysm vascular malform, aneurysm without, anicter neck, anicter sclera neck, anicter sclera op, anisocytoccasion, anomal, anomali, anomali hip, anomali hip stabl, anu sacral anomali, appear lesion, asymptomat, asymptomat given, asymptomat throughout, asymptomat without, coil aneurysm, cystic lesion, defect distal, dichorionicdiamniot twin, hypochromnorm anisocytoccasion, iliac aneurysm, intracrani aneurysm, lesion appear, lesion like, lesion seen, lesion seen within, like anemia, mm aneurysm, mycot aneurysm, myelo hypochromnorm anisocytnorm, saccular aneurysm, sacral anomali, sacral anomali hip, show aneurysm, small aneurysm, symptomat anemia, vascular anomali
perinatal period conditions	age preterm, complet done preliminari, done preliminari, done preliminari refer, due preterm, endometrium, gynecolog, immedi postop period, intrapartum, last menstrual, last menstrual period, menstrual, menstrual period, menstruat, mother present preterm, mother receiv intrapartum, normal preterm, normal preterm femal, normal preterm male, onset preterm, patient period, perinat, period anesthesia, period hypotens, period patient, periop period, postmenopaus, postop period, postop period patient, pregnanc complic cervic, preliminari, preliminari refer, preliminari report, preliminari report pfi, preliminari report wet, preliminari result, preliminarili, prematur week twin, preterm, preterm femal, preterm femal genitalia, preterm infant follow, preterm low, preterm male, preterm male test, stool pertin result, urinari symptom, urinari symptom past, utero, yearold woman prenat
injury and poisoning	accident, acid discharg medic, bleed like, bleed risk, bleed stabil, diff toxin, diff toxin assay, diffus process toxin, digoxin toxic, drug reaction, drug reaction attendinglast, drug toxic, due dehydr, get wound, given risk bleed, harm, harm suggest, health problem, high risk bleed, increas risk bleed, like bleed, medic side effect, per toxicolog, poison, process toxin, reaction patient, renal toxic, risk bleed, seen toxicolog, septic shock like, sign infect wound, stabil bleed, toxic, toxic ingest, toxic patient, toxicmetabol, toxicolog, toxicolog consult, toxicolog consult recommend, toxin, toxin assay, toxin assay final, toxin eia, toxin eia refer, toxin posit, toxin sent, toxin test, toxin test final, toxin testfin, toxin testfin inpati

external causes of injury	admiss trauma, back trauma, bleed due, brought trauma, complaint trauma, culprit lesion, diagnosi trauma, diagnosi traumat, due trauma, evid trauma, excess bend wound, exercis wound, extrem wound, facial trauma, fall trauma, follow trauma, followup trauma, head trauma, instruct follow trauma, lesion could, lesion origin, like traumat, lower extrem wound, pain due, pain like due, post trauma, post traumat, prior trauma, recent trauma, report wound, set trauma, side effect, sign trauma, sp trauma, stitl trauma, taken trauma, transfer trauma, trauma, trauma fall, trauma head, trauma histori, trauma histori present, trauma intens, trauma left, trauma neck, trauma right, trauma sp, trauma surgic, trauma surgic intens, traumat
supplementary	addendum, also, also consult, also consult given, also discuss, also followup, also review, also see, appendix, background suggest, comparison, comparison brief, complic brief, consid followup, consult, consult also, consult note, consult see, detail, detail brief, detail pleas refer, due contrast, extra, extravas note, final refer, followup also, followup new, given extra, given supplement, inform section, later note, manual, need also, need due complex, need supplement, note also, note followup, ref, section detail, section due, see addendum, see also, see detail, supplement, supplement given, supplement need, supplement recommend, supplementari, use supplement, without supplement