Data Evaluation and Enhancement for Quality Improvement of Machine Learning

Haihua Chen, Jiangping Chen, and Junhua Ding

***Abstract*—Poor data quality has a direct impact on the performance of the machine learning system that is built on the data. As a demonstrated effective approach for data quality improvement, transfer learning has been widely used to improve machine learning quality. However, the "quality improvement" brought by transfer learning was rarely rigorously validated, and some of the quality improvement results were misleading. This article first exposed the hidden quality problem in the datasets used to build a machine learning system for normalizing medical concepts in social media text. The system was claimed to have achieved the best performance compared to existing work on a machine learning task. However, the results of our experiments showed that the "best performance" was due to the poor quality of the datasets and the defective validation process. To address the data quality issue and build a high-performance medical concept normalization system, we developed a transfer-learning-based strategy for data quality enhancement and system performance improvement. The results of the experiments showed a strong correlation between the quality of the datasets and the performance of the machine learning system. The results also demonstrated that a rigorous evaluation of data quality is necessary for guiding the quality improvement of machine learning. Therefore, we propose a data quality evaluation framework that includes the quality criteria and their corresponding evaluation approaches. The data validation process, the performance improvement strategy, and the data quality evaluation framework discussed in this article can be used for machine learning researchers and practitioners to build high-performance machine learning systems. The code and datasets used in this research are available in GitHub (https://github.com/haihua0913/dataEvaluationML).**

***Index Terms*—Convolutional neural network (CNN), data quality, medical concept normalization, recurrent neural network (RNN), transfer learning.**

Manuscript received August 1, 2020; revised November 12, 2020 and March 23, 2021; accepted March 30, 2021. This work was supported in part by the National Science Foundation under Grant 1852249, in part by the National Security Agency under Grant H98230-20-1-0417, and in part by the State Key Laboratory for Novel Software Technology in Nanjing University, China, under Grant KFKT2019A19. Associate Editor: M. Nagappan. *(Corresponding author: Junhua Ding.)*

Haihua Chen and Jiangping Chen are with the Department of Information Science, University of North Texas, Denton, TX 76203 USA (e-mail: haihua.chen@unt.edu; jiangping.chen@unt.edu).

Junhua Ding is with the Department of Information Science, University of North Texas, Denton, TX 76203 USA, and also with the State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210093, China (e-mail: junhua.ding@unt.edu).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TR.2021.3070863.

Digital Object Identifier 10.1109/TR.2021.3070863

I. Introduction

DEEP learning [1] as the most crucial breakthrough in machine learning history has drawn great attention from academics as well as industries. Its success stories include winning the world champions in Go game playing [2], beating human experts on recognizing objects in images, and significantly improving the quality of voice recognition, natural language processing (NLP), and autopiloting. However, lacking of high-quality training data becomes a major threat to the usage of deep learning. Approaches, such as crowdsourcing web data [3] or transferring data from other domains [4], have been proposed for enhancing training data. Nevertheless, these data could introduce noises, such as invalid data and label noises. Although some carefully designed deep learning models are robust to massive label noises [5], the computing rule of "garbage in, garbage out" is still applicable to deep learning. An experimental study performed by Buolamwini and Gebru [6] has shown a face-recognition-based gender classification system that was implemented with machine learning algorithms produced 0.8% error rate for recognizing the faces of lighter skinned males, but as high as 34.7% error rate for recognizing the faces of darker-skinned females. The problem was due to the significant imbalance of the training datasets in skin colors. Rajpurkar *et al.* [7] reported a deep learning system called ChexNet that was developed for diagnosing pneumonia diseases based on chest X-ray images. They claimed that the CheXNet exceeded average radiologist performance on pneumonia detection on both sensitivity and specificity [7]. However, many radiologists and machine learning experts suspected the result due to the questionable dataset [8]. Eklund *et al.* [9] reported an investigation regarding the validity of several fMRI studies of weakly significant neuroimaging results that have been published. The problem is due to the inadequate validation of the statistical methods with real data in the fMRI studies [9]. Many experiments have shown that data with poor quality could negatively impact the performance of deep learning significantly [10], [11]. Other experiments have demonstrated that better quality of training data could improve the performance of deep learning [12], [13]. Therefore, a systematic evaluation of the quality of the dataset is critical for building a high-quality machine learning system. The evaluation result would offer guidelines for data enhancement and system performance improvement.

One of the purposes of this study is to understand the impact of data quality on the performance of machine learning systems that are built on the data. Data quality is a multidimensional

TABLE I
EXAMPLES OF MAPPING MEDICAL CONCEPTS TO TWITTER PHRASES

Medical Concept	Informal Phrase
Numbness	can't feel pain can't feel limbs lack of feeling
Influenza	fatigue
Depersonalization	make me feel different person

Note: The left column shows the formal medical concepts, whereas the right column shows the informal phrases.

concept on qualitative and quantitative attributes of data, and the definition varies under different contexts. For example, "data quality" could be about measuring the wrong data or missing data in the general context [14], or it means "satisfying the needs and preferences of its users" in a particular context [15], [16]. In this article, we define "data quality" as a measurement of data for fitting the purpose of building a machine learning system. Selected quality dimensions, such as comprehensiveness, correctness, and variety, will be used to assess the data quality [15], [17].

In this article, we first investigated an overclaimed performance improvement of a machine learning system. The problem was due to the poor quality of the datasets for building the system and the problematic validation process. Based on the investigation of the results, we proposed the data quality requirements to validate a machine learning system. Then, we experimented different strategies for performance improvement of the machine learning system that was built on the poor-quality datasets. Finally, we introduced a framework for evaluating the data quality to ensure the quality requirements of datasets for building a high-quality machine learning system. The data validation process, the performance improvement strategies, and the data quality framework are explained through studying a medical concept normalization system, which maps an informal medical term or phrase in social media text into a formal medical concept [18]. Table I lists some examples of medical concept normalization, where the informal phrase "fatigue" is mapped to the medical concept "influenza."

The medical concept normalization system we study in this article was first reported in [18]. It was built on deep learning models, which require a large amount of training data. The datasets presented in this article seem insufficient and there is a large portion of overlap between the training and the test data. However, the medical concept normalization system developed in [18] that did not introduce any new model or algorithm achieved much better performance comparing to other similar systems at that time. The performance achievement is suspicious and it might be due to the noises in the datasets [19]. Nevertheless, how can we find the problem? We developed a validation approach that evaluated the system with well-designed configurations of datasets and finally found the problem, which was caused by the overlap between the test datasets and the training datasets. The overlap of training and test datasets is not necessary to cause the inflation of performance as soon as the machine learning model is well trained and generalized. We will show when the overlap may produce problems and how the evaluation result can help the developer to improve the system under development in Section III.

The next question we will discuss is how we can improve the performance of the machine learning system that was developed on poor quality datasets. Different transfer-learning-based strategies, including fine-tuning and using advanced language models, were experimented for the performance improvement. Transfer learning [20] is a machine learning method that develops a new model through reusing and tuning a trained model. For example, one can build an image classification model for a specific image classification task by reusing and tuning AlexNet that was pretrained in ImageNet [21]. Transfer learning can also be used for improving the quality of a dataset, where a dataset for general applications is fine-tuned with a domain-specific dataset through transfer learning.

However, transfer learning is not always effective for performance improvement, as demonstrated in the experiments in Section IV. Only carefully selecting the source dataset and target dataset could produce the desired result. We designed a serial of experiments to explain the best practices for applying transfer learning for data quality improvement. Since data quality can greatly impact the performance of machine learning, it is necessary to define criteria for adequately evaluating it. We defined the three most crucial data quality criteria: comprehensiveness, correctness, and variety for evaluating the "fit for purpose" of datasets for machine learning, and proposed approaches for adequately evaluating the criteria. The three criteria are essential for building a high-performance machine learning system and guiding the application of transfer learning for quality improvement of machine learning.

In summary, this research will answer a few important questions regarding the data quality to the performance of machine learning.

1) Cross-validation (CV) might not be sufficient to validate the performance of a machine learning system since the test and the training datasets may share a significant amount of data if the original dataset contains many duplicated data items. In this case, can we design an approach for better evaluating the system? How can we use the evaluation result for improving the performance of the system under development? We conducted an experimental study to answer these questions and proposed the solutions based on the experiment results. The experiment results offer evidence to present the seriousness of the problem and show the necessity and the solution of assessing the quality of machine learning datasets.
2) If the quality of the dataset is not high enough for building a machine learning system with desirable performance, what are the best strategies that can be applied for improving the system performance? In particular, under what circumstances can transfer learning be used for performance improvement of the machine learning system? What are the best practices for using transfer learning to improve the performance of the machine learning system? We will answer these questions through experiments of a group of transfer-learning-based performance improvement strategies on the medical concept normalization system.

3) The data quality should be adequately evaluated before the data are used for building a machine learning system, and the evaluation results should be able to guide the system validation and performance improvement. We need to be clear about the criteria for adequately evaluating the data quality for machine learning. How can we measure the criteria and how the data quality evaluation result would help the quality improvement? This research will propose a data quality framework to answer these questions.

The remainder of this article is organized as follows. Section II introduces the research background on data quality, medical concept normalization, transfer learning, datasets, and experiment settings. Section III describes the validation of a medical concept normalization system to demonstrate the impact of data quality on the performance of the system. Section IV discusses the transfer-learning-based strategies for the quality improvement of the medical concept normalization system. Section V proposes a data quality framework for building a machine learning system. Section VI conducts a comprehensive literature review on medical concept normalization, data quality evaluation, and transfer learning. Finally, Section VII concludes this article.

II. Research Background

A. Data Quality

Data quality is on the comprehensive characterization and measurement of quantitative and qualitative properties of data. Wang and Strong defined data quality as "data that are fit for use by data consumers" [17]. Wand and Wang defined data quality as "the quality of mapping between a real-world state and an information system state" [22]. In this research, data quality means "satisfying the needs and preferences of its users or tasks" or the capability of the data for "fit for purpose" of building a machine learning system [15], [16]. It is about the characterization and measurement of the state of data for fitting the purpose of building a machine learning system. Selected quality dimensions, such as comprehensiveness, correctness, and variety, will be used for assessing the data quality [15], [17] for building a high-performance machine learning system.

B. Medical Concept Normalization

On social media platforms, such as Twitter, and online health forums, such as Ask a Patient,[1] users often share experiences and opinions on various health topics. They ask health-related questions, write reviews on medications, and describe the side effects they experienced while taking a drug. Information from these texts, if extracted and analyzed appropriately, can help users and professionals to better understand the medical and health problems and even to identify potential solutions. However, the lexical and grammatical variability of the language used in social media platforms poses a key challenge for extracting and analyzing the information. In particular, the frequent use of informal language, nonstandard grammar, and abbreviation forms and typos in social media texts contribute to the challenge. Medical concept normalization is a way to address the issue. It maps a user-generated text regarding a health or medical condition described in colloquial language to a medical concept in a standard ontology, such as Unified Medical Language System [23]. Table I lists some examples of medical concept normalization.

[1][Online]. Available: https://www.askapatient.com/

C. Medical Concept Normalization System

Limsopatham and Collier [18] proposed a machine learning approach for normalizing medical concepts in social media messages. The normalization was built on a neural-network-based text classification to learn the mapping between social media messages and medical concepts [18]. The normalization classifier was built on either a convolutional neural network (CNN) model [24] or a recurrent neural network (RNN) model [25]. The inputs of the CNN or RNN model are phrases and medical concepts that are embedded into vectors using a language model, such as Word2Vec [26]. They conducted the experiments on three datasets: TwADR-S, TwADR-L, and AskAPatient. The results of the experiments reported in [18] showed the CNN model with a language model, which was pretrained on Google News, achieved the best performance among all the three datasets (i.e., 0.4174, 0.4478, and 0.8141 regrading accuracy, respectively). However, the same experiments conducted by Lee *et al.* [19] showed that the RNN model outperformed the CNN model on both TwADR-L and AskAPatient datasets when the language model was trained with any of the seven different datasets, and the best performance (i.e., 0.2530 and 0.6504 regrading accuracy, respectively) is much lower than that was reported in [18]. The high accuracy might be overclaimed in [18] since the approaches applied in [18] and [19] were the same. The "performance improvement" reported in [18] might be due to the noises in the datasets, as pointed out by Lee *et al.* [19]. However, a systematic validation is needed to confirm the problem. We developed an approach for removing the noises in the dataset and conducted a systematic investigation. We found that the normalization performance decreased when the amount of noises decreased in the datasets.

D. Fine-Tuning for Performance Improvement

Fine-tuning [27] takes a machine learning model, which was trained for a task, to another task through fine-tuning the model with task-specific datasets. Fine-tuning is a specific transfer learning technique, which has been widely used in computer vision and NLP to build a task-specific system quickly and to improve the performance of the system [20]. In this research, the machine learning models for medical concept normalization were trained with datasets AskAPatient and TwADR-L. The words in the datasets are converted into word vectors as input to the model using a language model, such as Word2Vec. The quality of the language model would decide whether the semantics of a word are accurately captured in the corresponding vector. Therefore, the quality of the dataset is indirectly decided by the language model. Quality improvement of the language model would improve the quality of the dataset. The language

models, such as AWD-LSTM in ULMFit [28] or BERT [29], were trained on general text. We expect that fine-tuning the models with task-specific data would improve the performance of the machine learning system built on the data.

The general steps for fine-tuning a deep learning model are first to select a pretrained model and simply add layers in the model for implementing the target task, then freeze the model except for the least or the most significant layers and train the model using the target dataset. More layers might be unfrozen and the model is gradually trained until it is convergent.

E. Datasets

Several datasets are being used in our experiments. TwADR-L and AskAPatient are used for medical concept normalization, whereas datasets Cadec, Pubmed, Heathnews, and Big_tweet are used for fine-tuning language models. Combinations of these datasets are also developed for fine-tuning purpose.

1) Datasets for Medical Concept Normalization: Twitter message dataset TwADR-L and blog message dataset AskAPatient were first introduced by Limsopatham and Collier for medical concept normalization [18], and both are available for downloading in Zenodo.org [30]. The Twitter messages collected in TwADR-L are related to a set of drugs and adverse drug reactions (ADRs). It contains 1436 distinct twitter phrases, and each of them is mapped to one or more medical concepts of the 2220 medical concepts, which are defined in SIDER 4.1 drug profile databases.[2] This produces a total of 50 730 records in TwARD-L, and each record consists of one informal phrase and its corresponding medical concept. Among which, 48 057 records are used for training, 1256 records are used for validation, and 1427 are used for testing. However, only 273 among the 2220 medical concepts map to the Twitter phrases [18], which flags a comprehensiveness problem of the dataset. When a Twitter phrase is mapped to multiple concepts, the phrase is copied for multiple times so that the phrase appears multiple times in the dataset but associated with different concept each time. Therefore, the number of records is much higher than the number of phrases (as mentioned earlier, there are 1436 phrases and 50 730 records). Dataset AskAPatient includes 3749 phrases, and each of them is mapped to at least one of the 1036 medical concepts defined in SNOMED-CT and the Australian Medicines Terminology [18]. One phrase could be mapped to more than one medical concept, and multiple phrases could also be mapped to one medical concept. There are a total of 173 240 records in AskAPatient. Among which, 156 652 records are used for training, 7926 records are used for validation, and 8662 records are used for testing. Table II summarizes the two datasets. CV was used for the validation of the normalization system, and Table III summarizes the statistics of the records of the training dataset, validation dataset, and test dataset in each of the ten folds that were developed in [18].

However, each dataset includes more than 50% duplicated records (i.e., two records are exactly the same: the same phrase is mapped to the same medical concept) and has a significant amount of overlaps among the training datasets, their corresponding validation datasets, and test datasets [19]. An overlap existing in two datasets means the same record exists in the two datasets. For example, if a record in a training dataset is "Hunger – don't want to eat," and there is precisely the same record in a test dataset, then the record is considered as an overlapped record in the two datasets. Datasets AskAPatient and TwADR-L contain many overlapped records, as shown in Table IV, which describes the total overlaps, and the overlapped records in each of ten folds. From the table, one can easily find almost 100% of records in the validation dataset or the test dataset exist in the corresponding training dataset. In each fold, around 60% records in the test dataset exist in its corresponding training dataset. The duplicated records will not impact the performance of the

[2][Online]. Available: http://sideeffects.embl.de/

TABLE II
SUMMARY OF DATASETS TWADR-L AND ASKAPATIENT

Item	TwADR-L	AskAPatient
Medical Concepts	2,220	1,036
Phrases	1,436	3,749
Training Dataset	48,057	156,652
Validation Dataset	1,256	7,926
Test Dataset	1,427	8,662

Note: In the TwADR-L dataset, only 273 among 2220 medical concepts are mapped to the 1436 Twitter phrases. In the AskAPatient dataset, each of the 3749 phrases is mapped to at least one of the 1036 medical concepts.

TABLE III
SUMMARY OF THE NUMBER OF RECORDS IN EACH TEN-FOLD DATASET OF ASKAPATIENT AND TWADR-L

	AskAPatient			TwADR-L		
	Training	Valid.	Test	Training	Valid.	Test
Total	156652	7926	8662	48057	1256	1427
Fold 1	15612	845	867	4816	115	143
Fold 2	15631	826	867	4817	114	143
Fold 3	15700	758	866	4791	140	143
Fold 4	15672	786	866	4812	119	143
Fold 5	15630	828	866	4811	120	143
Fold 6	15675	783	866	4801	130	143
Fold 7	15710	748	869	4819	112	143
Fold 8	15659	799	866	4790	142	142
Fold 9	15647	811	866	4788	144	142
Fold 10	15716	742	866	4812	120	142

Note: The ten-fold dataset is generated by Limsopatham and Collier [18]. The first row lists the total number of records from all folds.

TABLE IV
SUMMARY OF THE NUMBER OF OVERLAPPED RECORDS IN THE DATASETS

	AskAPatient		TwADR-L	
	Tr. ∩ Valid.	Tr. ∩ Test	Tr. ∩ Valid.	Tr. ∩ Test
Total	7926	8659	1256	1427
Fold 1	501	523	61	87
Fold 2	483	522	77	80
Fold 3	480	507	77	88
Fold 4	464	507	81	85
Fold 5	516	524	65	96
Fold 6	468	543	71	93
Fold 7	448	511	66	81
Fold 8	468	532	82	87
Fold 9	497	527	81	80
Fold 10	494	528	74	86

Note: In each fold, around 60% records in the test dataset exist in its corresponding training dataset.

machine learning system if they are directly extracted from the original data sources. However, the large amount of overlaps between the training and test data could produce misleading validation results when the datasets are used for testing the system.

2) Datasets for Transfer Learning and Fine-Tuning: The quality of datasets TwADR-L and AskAPatient is fairly low. For example, dataset TwADR-L only includes a few formal medical concepts, and both datasets are too small to be considered as complete or comprehensive. Therefore, it is necessary to identify new datasets to fine-tune the machine learning models in order to improve the performance of the machine learning system. We collected following four datasets from different sources that are closely related to medical concept normalization to fine-tune the medical concept normalization system that was developed on datasets TwADR-L and AskAPatient.

1) Cadec dataset (abbreviated with "Cadec"): This dataset represents CSIRO adverse drug event (ADE) corpus (Cadec). It contains annotated corpus of medical forum posts on patient-reported ADEs. It is useful for studies in the area of information extraction, or more generically text mining, from social media to detect possible ADRs. We collected 7000 posts that are related to 12 drugs: *arthrotec, cambia, cataflam, diclofenac potassium, diclofenac sodium, flector, liptor, pennsaid, solarize, voltaren, voltaren-XR*, and *zipsor*.
2) Pubmed dataset (abbreviated with "Pub"): It includes 7000 sentences extracted from abstracts related to same 12 drugs mentioned in Cadec dataset from Pubmed database.
3) Healthnews dataset (abbreviated with "Health"): It includes 7000 news extracted from 16 resources: *bbchealth, cbchealth, cnnhealth, everydayhealth, foxnewshealth, gdnhealthcare, goodhealth, kaiserhealthnews, latimeshealth, msnhealthnews, nbchealth, nprhealth, nytimeshealth, reutershealth, usnewshealth*, and *wsjhealth*.
4) Big_tweet dataset (abbreviated with "Big_t"): It includes 7000 tweet sentences that were extracted from Healthcare Twitter analysis project [31], which contains over six million tweets concerning a wide range of medical conditions for six months in 2014.

The datasets Pubmed and Healthnews contain many formal medical concepts, whereas datasets Cadec and Big_tweet include only few formal medical concepts. Therefore, it would be interesting to see how the combinations of two types of datasets would impact the performance of transfer learning.

F. Language Models

Pretrained language models, as an essential component of modern NLP, can offer significant performance improvement over embeddings learned from scratch through fine-tuning [32]. ULMFit [28] is a well-known fine-tuning technique that has been successfully used for the implementation of different text classification tasks. Its fine-tuning is applied to the language model AWD-LSTM. BERT [32] is the most widely adopted pretrained language model that can be fine-tuned for specific tasks. Our fine-tuning experiments will be developed on ULMFit and BERT.

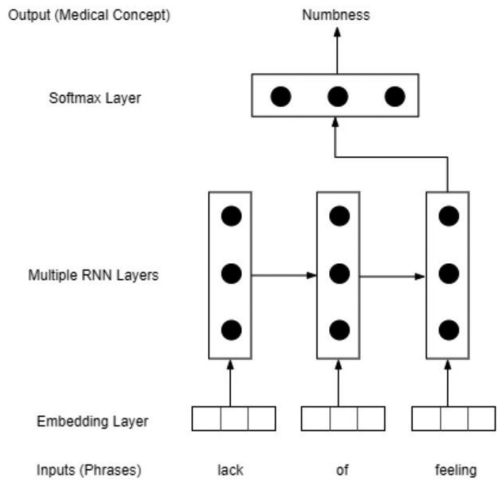

Fig. 1. Architecture of the RNN model for medical concept normalization [18].

ULMFit trains a general domain language model AWD-LSTM with the open dataset WiKiText-103 [33], which is the WiKiText long-term dependence language modeling dataset. It includes 100 million tokens extracted from the set of verified good and featured articles on Wikipedia. The dataset is well suitable for training models that work for long-term dependencies, such as the models used in this article. There are two pretrained language models in BERT: BERT-base and BERT large, which are trained on the same data—the entire Wikipedia with 2500 million words and Book Corpus with 800 million words, and tasks with different numbers of parameters [32]. In this article, we use the BERT-base model, which contains a 12-layer bidirectional transformer encoder block with hidden size 768 and 12 heads.[3] BERT takes an input of a sequence of no more than 512 tokens and outputs the representation of the sequence [32].

G. Deep Learning Models

In this research, RNN models were used for normalizing medical concepts since the RNN model outperformed the CNN model, as being proven by Lee *et al.* in [19]. As shown in Fig. 1, the input to the RNN model is word embeddings of phrases, and the output from the Softmax layer is the corresponding medical concepts of the input phrases. The word embeddings are produced with a language model, such as Word2Vec [26], GloVe [34], or BERT [29]. For example, the language model used in [18] was built on Word2Vec that was trained with a dataset that includes 100 billion words from Google News and 854 million words of medical articles [18]. The dimension of the embedding vector is 300, which represents 300 closest words to the current word. The RNN model consists of multiple recurrent layers, and the output from the last layer is used for mapping to

[3][Online]. Available: https://storage.googleapis.com/bert_models/2018_10_18/uncased_L-12_H-768_A-12.zip

TABLE V
EXPERIMENT FRAMEWORK

MCN Datasets	Language Model	Fine-tuning Datasets
AskAPatient	ULMFit	no-tune
TwADR-L	ULMFit	no-tune
AskAPatient	ULMFit	Ask
TwADR-L	ULMFit	Twadr
AskAPatient, TwADR-L	ULMFit	Pub
AskAPatient, TwADR-L	ULMFit	Cadec
AskAPatient, TwADR-L	ULMFit	Health
AskAPatient, TwADR-L	ULMFit	Big_t
AskAPatient, TwADR-L	ULMFit	Pub + Cadec
AskAPatient, TwADR-L	ULMFit	Pub + Big_t
AskAPatient	ULMFit	Pub + Ask
AskAPatient	ULMFit	Cadec + Ask
AskAPatient	ULMFit	Health + Ask
AskAPatient	ULMFit	Big_t + Ask
AskAPatient	ULMFit	Pub + Cadec + Ask
AskAPatient	ULMFit	Pub + Big_t + Ask
TwADR-L	ULMFit	Pub + Twadr
TwADR-L	ULMFit	Cadec + Twadr
TwADR-L	ULMFit	Health + Twadr
TwADR-L	ULMFit	Big_t + Twadr
TwADR-L	ULMFit	Pub + Cadec + Twadr
TwADR-L	ULMFit	Pub + Big_t + Twadr
AskAPatient, TwADR-L	BERT	

Note: The trained models were also tested with different test datasets.

a medical concept with a Softmax function, as shown in Fig. 1. The details of the models can be found in the papers [18], [19].

H. Experiment Setting

Based on the aforementioned datasets, pretrained language models, and deep learning models, we designed a group of experiments to investigate the research questions. The settings of these experiments are summarized in Table V. The objectives of these experiments are described as follows.

1) Experiments to investigate the overclaimed performance issue of the medical concept normalization system. The experiment results will guide us to develop a data-driven approach for better validating a machine learning system and understanding how data quality affects the performance of a machine learning system.
2) Experiments the strategies for performance improvement of the medical concept normalization systems that were developed on datasets TwADR-L and AskAPatient. The experiments were built on the fine-tuning technique. ULMFit was first fine-tuned with datasets TwADR-L and AskAPatient for the medical concept normalization to investigate the effectiveness of fine-tuning to the performance improvement.
3) Experiments of fine-tuning the medical concept normalization systems with more datasets to understand how the target dataset would impact the performance of machine learning. We compared the performance of the ULMFit-based medical concept normalization models that were fine-tuned with the datasets built on datasets Cadec, Pubmed, Healthnews, Big_tweet, TwADR-L, and AskAPatient. The medical concept normalization classifier was trained and validated with datasets TwADR-L and AskAPatient. In order to investigate the impact of the capability of language models on the performance of machine learning, BERT was also fine-tuned with target-task specific datasets TwADR-L and AskAPatient for the medical concept normalization. The experiment settings and results will be reported in Section IV.

III. SYSTEM VALIDATION

When checking datasets AskAPatient and TwADR-L, we found there were a large amount of overlaps existing in different datasets. Lee *et al.* also pointed out that the overclaimed performance improvement could be due to the large overlap between the training and the test data [19]. However, the overlap is not necessary to contribute to the inflation of the normalization performance. If the RNN model was well trained and generalized, the overlap would not cause the problem. However, many researchers might be too confident with their machine learning models and the training process. As soon as they see the validation loss is stable and lower than a threshold, the training is stopped and the model is released. Nevertheless, the lower validation loss could be due to the overfitting caused by the overlap existing in the training, validation, and test datasets.

To our knowledge of deep learning, we have not seen a systematic study to investigate the impact of the overlap in datasets on the performance of deep learning. The widely used *N*-fold CV might not be sufficient to reduce the overlap since a dataset, such as TwADR-L, contains many duplicates. We therefore conducted a series of experiments to identify the change of the accuracy of normalizing medical concepts in datasets AskAPatient and TwADR-L with a different portion of overlapped data. The medical concept normalization system is built on the RNN model proposed in [18] and language model AWD-LSTM trained with WikiText-103 corpus [35].

A. Validation Results

In order to build the new test datasets with different percentage of overlapped data between the training and test datasets, we first generated a file that contains only the unique records in the datasets. We select one fold as the test dataset from the ten folds of the dataset and leave the other nine folds as training and validation datasets. If a record in the test dataset also exists in the training data or validation dataset, then the record and its duplicates are removed from the test dataset. Then, we search for records that do not exist in the testing dataset from the training or validation datasets based on the file that contains only unique records and add some of these records into the test dataset to ensure it contains 10% records of the entire dataset. The records and their duplicates that were added from the training or validation datasets to the test dataset were removed from their original datasets. Using this way, we can create a test dataset that does not contain any overlapped record. We then remove 10% records from the test dataset that we just tested and then add 10% of records that exist in the training dataset. Therefore, 10% of the test dataset are overlapped with the data in the training dataset. Using the same schema, 20%, 30%,..., and 100% overlapped records were tested one by one. If the RNN model was well trained, the accuracy of the normalization of the medical concepts with different test datasets should be almost

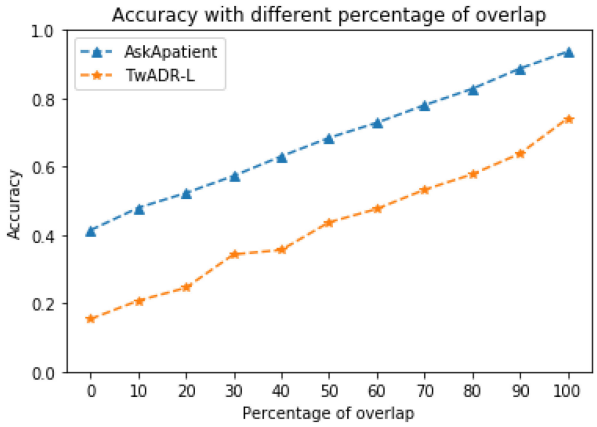

Fig. 2. Accuracy of normalizing medical concepts on datasets AskAPatient and TwADR-L with different percentages of overlapped records in the test dataset.

TABLE VI
ACCURACY OF MEDICAL CONCEPT NORMALIZATION ON DATASETS ASKAPATIENT AND TWADR-L WITH A DIFFERENT PERCENTAGE OF OVERLAPPED RECORDS IN THE TEST DATASET

Overlapped	AskAPatient	TwADR-L
0%	0.4138	0.1538
10%	0.4794	0.2076
20%	0.5229	0.2453
30%	0.5729	0.3431
40%	0.6303	0.3561
50%	0.6844	0.4371
60%	0.7285	0.4762
70%	0.7801	0.5325
80%	0.8279	0.5776
90%	0.8874	0.6381
100%	0.9366	0.7413

Note: The accuracy increases proportionally with the increment of the percentage of the overlapped data in the test dataset, as can also be seen in Fig. 2.

the same even overlapped data exist. However, we found that the accuracy increases proportionally with the increment of the percentage of the overlapped data in the test dataset.

The results of the experiment show that the RNN model was not well generalized. The claimed high accuracy was obtained by applying the problematic test datasets. The same experiment was conducted on both datasets, and the results support the same conclusion. Therefore, it is necessary to develop a strategy to enhance the data for improving the performance of the machine learning system.

The difference in the accuracy of the medical concept normalization among the ten folds is ignorable. Therefore, we explain our experiments for evaluating the datasets using only one fold in the following sections. Fig. 2 shows the testing results of the accuracy of medical concept normalization on datasets AskAPatient and TwADR-L with different percentage of overlapped data, and Table VI lists the result.

B. Discussion

It is well understood that a large overlap between training and test data could contribute to the nonexist high performance of a machine learning model [19]. However, how do we know whether the overlap causes a problem since the overlap does exist more or less and it is not necessary to cause the problem? In this research, we designed a validation approach that tests the model with a different percentage of the overlap between the training and test datasets. Suppose the percentage of the overlap increases, the performance of the model under test increases accordingly, the model might not be well trained. The test result could help developers to take further actions for the system improvement.

The results also help us to investigate another related question: when is the test dataset adequate for testing a machine learning model? Test coverage criteria are used to measure the testing adequacy in software testing. The coverage criteria are normally defined on the code and functions of the software under test. The coverage criteria cannot be directly used for testing a machine learning model since the logic of a machine learning model is not explicitly predefined but learned from data. Therefore, a test coverage criterion should be defined with the data to measure the test adequacy. We already showed that the test data should include a large amount of new data, but the requirement is not sufficient since the new data could only represent a small fraction of the population the data collected from. An ideal criterion should be defined on the representativeness of the test data to the population, which can be measured by comparing the distributions of the test data and the population in selected features. We will discuss the data quality evaluation in detail in Section V.

IV. QUALITY IMPROVEMENT USING TRANSFER LEARNING

The machine learning models for the medical concept normalization in this research were trained with datasets AskAPatient and TwADR-L. The phrases and the medical concepts in the datasets are embedded as word vectors using a language model. The vectors are input to a machine learning model, such as a classifier that implements the mapping from an informal phrase to a medical concept. Therefore, the data quality is also indirectly decided by the quality of the language model since it converts phrases into vectors. A language model decides whether the semantics of phrases and medical concepts are accurately captured in the word vectors. However, the language models, such as ULMFit and BERT, were trained on general text. The actual language model of ULMFit is AWD-LSTM, but the model of ULMFit is the same as AWD-LSTM except for the classification layer. Therefore, we use ULMFit to name its language model when it does not cause confusion. In order to improve the quality of word embedding, transfer learning, in particular, the fine-tuning is used to fine-tune the language model. In this section, we discuss the approaches for improving the quality of machine learning through fine-tuning the language model and its normalization task jointly and the language model alone. The purpose is to investigate how the dataset used for fine-tuning would impact the performance of fine-tuning and then further impact the performance of machine learning. Based on the investigation, we summarize the best practices for using fine-tuning for performance improvement of a machine learning system.

A. Fine-Tuning Medical Concept Normalization Model

We fine-tuned the medical concept normalization model by following the procedure developed in ULMFit [28], which includes two phases: first, fine-tune the language model using task-specific datasets, and second, fine-tune the entire medical concept normalization model that combines the language model and the classifier using the training dataset. The language model of ULMFit includes an embedding layer, followed by three layers of LSTM [36], and a Softmax layer for the classification as the last layer [28]. The language model was pretrained with general text. We fine-tuned it with a dataset that is related to medical concept normalization. In this way, we hope the fine-tuned model would better represent the medical concept normalization language. After the language model is fine-tuned, two additional layers implementing the classifier are added to the model with one layer with rectified linear unit (ReLU) activations, and followed by the last layer with Softmax activations. The ReLU is fully connected with the output of the language model, and its output is fully connected to the Softmax layer, which carries out the normalization task. The layers of the language model are "unfrozen" gradually during fine-tuning to ensure the knowledge learned from the vast corpus is mostly kept. ULMfit uses a discriminative fine-tuning technique that uses different learning rate for each layer to effectively tune the model [28]. The fine-tuning procedure of ULMFit is summarized as follows.

1) Fine-tune the general language model with datasets AskAPatient or TwADR-L using the discriminative fine-tuning technique, which uses the slanted triangular learning rates to calculate the learning rate for each layer. The slanted triangular learning rate first linearly increases the learning rate and then linearly decays it [28].
2) Use the fine-tuned language model as a base, add a classifier to the model for the medical concept normalization task, which includes a ReLU layer and a Softmax layer that outputs a probability distribution over medical concepts.
3) Fine-tune the entire model (i.e., the language model and the classifier together) using a gradual unfreezing technique with the training dataset in AskAPatient or TwADR-L. It first freezes all layers of the model except the first layer (we consider the Softmax as the last layer), which is fine-tuned with one epoch. Then, the second layer is unfrozen and fine-tuned, and the step is repeated until all layers are unfrozen and tuned [28].

B. Initial Fine-Tuning Experiments

The ULMFit model, including the general language model and the classifier together, that was designed for general text classification tasks can be fine-tuned for the medical concept normalization. The model is also an RNN model, same as the model we used in Section III so that it can be directly used for building the medical concept normalization. We conducted four different experiments, which are as follows.

1) Train ULMFit model with the training dataset of AskAPatient, and test the trained model using the test dataset. Fine-tuning was not implemented in this experiment.
2) Fine-tune ULMFit language model using dataset AskAPatient, and then fine-tune the entire ULMFit model using the training dataset of AskApatient, and test the trained model using the test dataset of AskAPatient.
3) Train ULMFit model with the training dataset of TwADR-L, and test the trained model using the test dataset. Fine-tuning was not implemented in this experiment.
4) Fine-tune ULMFit language model using dataset TwADR-L, and then fine-tune the entire ULMFit model using the training dataset of TwADR-L, and test the trained model using the test dataset of TwADR-L.

TABLE VII
SUMMARY OF THE CV RESULTS OF MEDICAL CONCEPT NORMALIZATION THAT WAS FINE-TUNED OR WAS NOT FINE-TUNED ON DATASETS ASKAPATIENT OR TWADR-L

A(Tuned)	A(No Tuned)	T(Tuned)	T(No Tuned)
0.7630	0.6610	0.3158	0.2670
0.7455	0.6626	0.3893	0.2746
0.7834	0.6626	0.3741	0.2835
0.7490	0.6604	0.4322	0.3520
0.7836	0.6473	0.3445	0.3040
0.7558	0.5731	0.3101	0.2875
0.7898	0.6594	0.4144	0.3473
0.7706	0.6596	0.4255	0.3442
0.7741	0.6627	0.3497	0.2971
0.8010	0.6586	0.4202	0.3900
0.7716	0.6507	0.3776	0.3147

Note: Columns one and three show fine-tuning the model (with ULMFit) can increase the accuracy of medical concepts normalization on both datasets. The last row shows the accuracy performance averaged across the ten folds.

The experiments were conducted on an Ubuntu 18.04.3 LTS machine with 1 NVIDIA Tesla Titan V GPU, 8 Intel(R) CPUs (i7-9700 @3.00 GHz), and 128 GB of RAM. The embedding dimension is 400. The batch size and epoch were dynamically adjusted during experiments. Most hyperparameters were the same as those used in pretraining [28]: Adam with $\beta_1 = 0.7$ and $\beta_2 = 0.999$, weight dropout of 0.5 to the RNN hidden-to-hidden matrix, learning rates were dynamically changed for fine-tuning different layers.

CVs of the normalization of the medical concepts on datasets AskAPatient and TwADR-L were conducted. Table VII summarizes the CV results of the medical concept normalization on datasets AskAPatient and TwADR-L. In the table, A(tuned) and A (no tuned) represent the ULMFit model was and was not fine-tuned using AskAPatient, respectively; T(tuned) and T(no tuned) represent the ULMFit model was and was not fine-tuned using TwADR-L, respectively. Figs. 3 and 4 show a comparison of the medical concept normalization accuracy using CV on datasets AskAPatient and TwADR-L, respectively.

The results in Table VII, Figs. 3 and 4 show that fine-tuning the model can increase the accuracy of medical concept normalization on both datasets. It implies that the performance of a model that is trained on poor quality datasets can be improved using well designed fine-tuning.

C. Fine-Tuning the Model With More Datasets

The initial fine-tuning results of the medical concept normalization showed the potential of transfer learning for improving

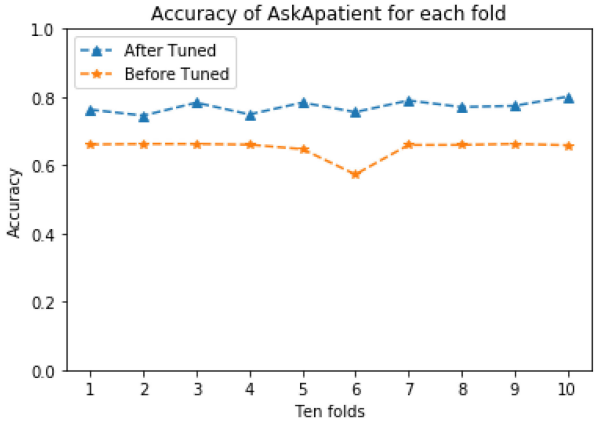

Fig. 3. Comparison of the accuracy of medical concept normalization that was fine-tuned or was not fine-tuned on dataset AskAPatient.

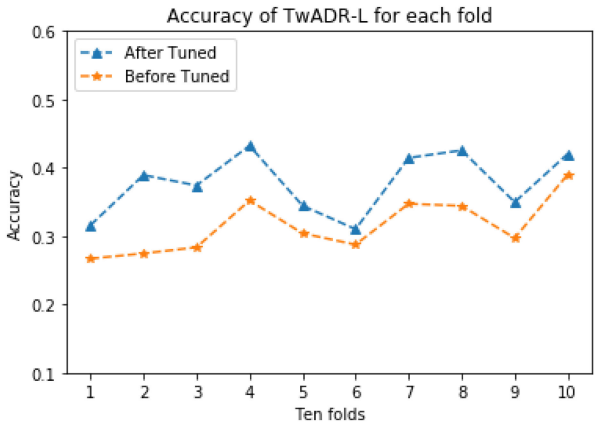

Fig. 4. Comparison of the accuracy of medical concept normalization that was fine-tuned or was not fine-tuned on dataset TwADR-L.

the quality of machine learning. However, the datasets still suffer from an overlap problem, and the quality of the datasets is essentially low. For example, dataset TwADR-L only includes a few formal medical concepts, and it is too small. We would like to know whether the performance of the medical concept normalization can be further improved through fine-tuning the models with new datasets.

We collected four datasets: *Pubmed*, *Healthnews*, *Cadec*, and *Big_tweet*, as introduced in Section II and created the combinations of these datasets. We conducted the experiments using these datasets to fine-tune the ULMFit model by following the same process discussed in Section IV-A. The training, validation, and test datasets are still the same as the original ones from datasets AskAPatient and TwADR-L, and the training and testing processes are the same as those used for the experiments of the medical concept normalization described in previous sections. The environment and parameter settings of the experiment for fine-tuning ULMFit are the same as those used in Section IV-B.

Table VIII lists the CV accuracy of the medical concept normalization on the model ULMFit and BERT. The language model of ULMFit was fine-tuned using different datasets and BERT was fine-tuned using the datasets AskAPatient or TwADR-L. All models were trained and tested using the training datasets and the testing datasets in dataset AskAPatient. The first row of the table lists the language models used in the experiments, and the second row lists the datasets used for fine-tuning the language model. For example, Pub means the language model was fine-tuned using dataset Pubmed, Pub+Cadec means the combination of datasets Pubmed and Cadec, Pub+Ask means the combination of datasets Pubmed and AskAPatient, and all others follow the same format. The third row to the 13th row represent the result at each fold of the CV results, and the last row is the average of the CV results across the ten folds. Table IX has the same format, but the models were trained and tested using the training datasets and the testing datasets in dataset TwADR-L.

Tables X and XI list the results of the same experiments as those shown in Tables VIII and IX, respectively, except the validation was conducted on test datasets that have different percentage of overlapped records from those in the training dataset.

D. Fine-Tuning Language Model

Since BERT is a very powerful state-of-the-art language model, we would like to know how the choice of a language model would impact the performance of the machine learning model. BERT was pretrained with a huge amount of general text and it can be fine-tuned with task-specific data to achieve the best result [29]. We fine-tuned ULMFit in two phases: fine-tune the language model and fine-tune the classifier together with the language model. However, we only fine-tune BERT language model. The output of the final transformer layer of the BERT language model is then used as the feature sequences to be fed to the classifier for the medical concept normalization.

We conducted the fine-tuning of language model BERT with the task-specific data AskAPatient or TwADR-L for medical concept normalization. The training, validation, and test datasets are still the same as the original ones from dataset AskAPatient or TwADR-L, and the training and testing processes are the same as those used for the experiments of medical concept normalization described in the previous sections. In order to comprehensively compare the performance with the experiments of fine-tuning the ULMFit model, we also conducted experiments with test datasets that were designed with different percentages of overlapped data. The language model is BERT-Base-Uncased, which has been introduced in Section II-F. We fine-tuned the BERT model on 1-T TITAN V GPU and set the batch size to 16, with a max sequence length of 128 and a learning rate of 2e-5 to ensure that the GPU memory is fully utilized. The dropout probability is always kept at 0.1. We use Adam with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. We empirically set the max number of the epoch and saved the best model on the validation set for testing.

The last columns in Tables VIII and IX are the results of the accuracy of the medical concept normalization on datasets AskAPatient and TwADR-L, respectively, by using BERT-based normalization model, whose language model is fine tuned with the informal phrases in the AskAPatient or TwADR-L dataset. The classifier for medical concept normalization is trained and

TABLE VIII
SUMMARY OF THE CV RESULTS OF MEDICAL CONCEPT NORMALIZATION ON DATASETS ASKAPATIENT USING FINE-TUNED ULMFIT AND BERT

ULMFit												BERT
Pub	Pub+ Cadec	Pub+ Big_t	Pub+ Ask	Pub+Cadec +Ask	Pub+Big_t +Ask	Cadec	Cadec+ Ask	Big_t	Big_t+ Ask	Health	Health+ +Ask	
0.5502	0.6678	0.5629	0.7278	0.7682	0.7290	0.6159	0.7428	0.2318	0.7255	0.2768	0.7370	0.8526
0.5178	0.6551	0.5433	0.7728	0.7647	0.7693	0.6332	0.7543	0.2053	0.7728	0.2745	0.7762	0.8365
0.5139	0.6755	0.5358	0.7379	0.7834	0.7506	0.5993	0.7760	0.2494	0.7644	0.2494	0.7159	0.8474
0.5370	0.6697	0.5427	0.7621	0.7656	0.7148	0.5947	0.7621	0.2148	0.7113	0.2483	0.7055	0.8181
0.5346	0.6986	0.5704	0.7818	0.7829	0.7402	0.5866	0.7356	0.2379	0.7263	0.2598	0.7471	0.8558
0.5185	0.6686	0.5612	0.7921	0.7794	0.7725	0.5947	0.7737	0.2448	0.7714	0.2806	0.7887	0.8427
0.5370	0.6524	0.5508	0.7540	0.7610	0.7286	0.6686	0.7829	0.2344	0.7171	0.2471	0.7749	0.8684
0.5219	0.6897	0.5612	0.6945	0.7806	0.7702	0.6016	0.6940	0.1871	0.7575	0.2864	0.6952	0.8437
0.5439	0.6597	0.5266	0.7783	0.7941	0.7598	0.6120	0.7691	0.2252	0.7148	0.2402	0.7425	0.8595
0.5219	0.6882	0.5624	0.7714	0.8369	0.7667	0.6524	0.7610	0.2344	0.7379	0.2748	0.7748	0.8665
0.5294	0.6725	0.5517	0.7573	0.7817	0.7502	0.6159	0.7552	0.2265	0.7399	0.2638	0.7458	0.8491

Note: BERT achieved the best performance. As for ULMFit fine-tuning, the performance varied using different datasets for fine-tuning. However, the combinations of Pub, Cadec, and Ask achieved a higher accuracy. The last row shows the accuracy performance averaged across the ten folds.

TABLE IX
SUMMARY OF THE CV RESULTS OF MEDICAL CONCEPT NORMALIZATION ON DATASETS TWADR-L USING FINE-TUNED ULMFIT AND BERT

ULMFit												BERT
Pub	Pub+ Cadec	Pub+ Big_t	Pub+ Twadr	Pub+Cadec +Twadr	Pub+Big_t +Twadr	Cadec	Cadec+ Twadr	Big_t	Big_t+ Twadr	Health	Health+ +Twadr	
0.2782	0.3560	0.3846	0.4895	0.4615	0.4685	0.4196	0.4755	0.2869	0.4476	0.2608	0.4196	0.4056
0.2719	0.3771	0.2727	0.3497	0.3636	0.3706	0.3147	0.3706	0.2719	0.3077	0.2807	0.3497	0.5000
0.2357	0.3357	0.2727	0.3916	0.4056	0.4195	0.3916	0.3916	0.2285	0.3846	0.2642	0.3916	0.3472
0.2857	0.3865	0.2937	0.3427	0.3986	0.2797	0.3287	0.3636	0.2941	0.3217	0.2689	0.3147	0.4631
0.2500	0.3583	0.3007	0.4126	0.3916	0.3636	0.3636	0.4196	0.2250	0.4266	0.2333	0.4266	0.4417
0.3153	0.3538	0.2867	0.4266	0.3916	0.4266	0.3916	0.4685	0.2461	0.4126	0.2615	0.4266	0.2868
0.2678	0.3928	0.2867	0.3636	0.4126	0.3566	0.3916	0.3706	0.2678	0.3636	0.2321	0.3986	0.3929
0.2957	0.3873	0.3028	0.3592	0.3803	0.3380	0.3239	0.4155	0.2887	0.3310	0.2957	0.3873	0.4606
0.2847	0.3451	0.2676	0.3451	0.3380	0.3169	0.3099	0.3521	0.2916	0.3239	0.2638	0.3380	0.3333
0.3250	0.4250	0.2887	0.4155	0.4417	0.3943	0.2958	0.3521	0.3500	0.3873	0.3250	0.3873	0.4500
0.2810	0.3718	0.2957	0.3896	0.3985	0.3734	0.3531	0.3979	0.2751	0.3707	0.2686	0.3840	0.4171

Note: BERT achieved the best performance. As for ULMFit fine-tuning, the performance varied using different datasets for fine-tuning. However, the combinations of Pub, Cadec, and Twadr achieved a higher accuracy. The last row shows the accuracy performance averaged across the ten folds.

TABLE X
VALIDATION RESULTS OF MEDICAL CONCEPT NORMALIZATION ON DATASET ASKAPATIENT WITH A DIFFERENT PERCENTAGE OF OVERLAPPED DATA

	ULMFit									BERT
Overlapped	Cadec	Pub+ Cadec	Cadec+ Ask	Pub+Cadec +Ask	Pub+ Big_t	Pub+ Ask	Pub+Big_t +Ask	Health+ Ask	Big_t+ Ask	
0%	0.3738	0.3770	0.4311	0.4311	0.3410	0.4066	0.4098	0.4475	0.4311	0.6104
10%	0.3891	0.4156	0.4347	0.4626	0.3803	0.4905	0.4435	0.4611	0.4347	0.6134
20%	0.4346	0.4332	0.5118	0.5131	0.3926	0.5249	0.5039	0.5301	0.5118	0.6654
30%	0.4897	0.5039	0.5538	0.5745	0.4243	0.5768	0.5562	0.5367	0.5538	0.6927
40%	0.4980	0.5118	0.5972	0.5914	0.4499	0.6189	0.5756	0.6012	0.5972	0.7520

Note: Similar patterns found as Tables VI and VIII.

tested using datasets in AskAPatient and TwADR-L, respectively. The last columns in Tables X and XI are the results of the same trained models, but they were tested with different percentage of overlapped data in datasets AskAPatient and TwADR-L, respectively.

E. Discussion

From the results of fine-tuning experiments described earlier, we summarize the insights as follows.

1) When the model ULMFit was fine-tuned with a dataset, such as Pubmed, Big_tweet, and Pubmed+Cadec, but the dataset did not include the target task-specific dataset AskApatient or TwADR-L, the accuracy of the normalization on the same test datasets was not improved compared to the experiments that did not fine-tune the model. As a matter of fact, the accuracy scores were even lower in most of the cases compared to the original experiments that did not fine-tune the model. For example, the average accuracy of original experiments on AskAPatient is around 65.07%, as shown in Table VII. However, the fine-tuned results with these datasets are from around 22.65% on dataset Big-tweet to 67.25% on dataset Pubmed+Cadec (it is the only one that is slightly higher than the original one), as shown in Table VIII.

The results of the experiments on dataset TwADR-L also has the similar pattern: the average accuracy of the original

TABLE XI
VALIDATION RESULTS OF MEDICAL CONCEPT NORMALIZATION ON DATASET TWADR-L WITH A DIFFERENT PERCENTAGE OF OVERLAPPED DATA

	ULMFit									**BERT**
Overlapped	Cadec	Pub+ Cadec	Cadec+ Twadr	Pub+Cade +Twadr	Pub+ Big_t	Pub+ Twadr	Pub+Big_t +Twadr	Health+ Twadr	Big_t+ Twadr	
0%	0.0887	0.1598	0.1667	0.1806	0.1319	0.1944	0.1528	0.1736	0.1667	0.1319
10%	0.1500	0.1750	0.2063	0.2063	0.1625	0.2125	0.2188	0.2000	0.2125	0.2222
20%	0.2277	0.2500	0.2889	0.3056	0.2389	0.2889	0.3000	0.2722	0.2722	0.2411
30%	0.2390	0.2878	0.3366	0.3610	0.2634	0.3610	0.3463	0.3317	0.3073	0.2813
40%	0.2783	0.3515	0.4017	0.3975	0.3180	0.3975	0.4017	0.4142	0.4017	0.3370

Note: Similar patterns found as Tables VI and IX.

experiments is around 31.47%, as shown in Tables VII, but the fine-tuned results with these datasets are from 26.86% on dataset Healthnews to 37.18% on dataset Pubmed+Cadec, as shown in Table IX. The result in dataset Pubmed+Cadec is fairly higher than the original one, and another one that is higher than the original one is on dataset Cadec, which is 35.31%.

2) When the model ULMFit was fine-tuned with different datasets, the accuracy of the normalization could be significantly different. When multiple datasets were combined together to fine-tune the model, it could achieve better normalization accuracy compared to only using the individual dataset. For example, when the ULMFit was fine-tuned with dataset Pubmed or Cadec, the average accuracy on dataset AskAPatient is 52.94% and 61.59%, respectively, and the accuracy is increased to 67.25% (i.e., at least 9.2% increase of the accuracy) when the model was fine-tuned with the dataset Pubmed+Cadec that combines both Pubmed and Cadec, as shown in Table VIII. The same pattern also shows in the results of the model that was tested with dataset TwADR-L, as shown in Table IX. However, not all combined datasets achieved the same accuracy. For example, when the model was fine-tuned with the combined dataset Pubmed+Big_tweet, the average accuracy is only 55.17%, as shown in Table VIII. The same pattern also exists in Table IX.
3) When the model ULMFit was fine-tuned with a combination dataset that includes the target task-specific dataset AskAPatient or TwADR-L, the accuracy of the normalization could be significantly increased compared to the dataset that does not include the target task-specific dataset. The accuracy is near to or higher than the accuracy of the model when it was fine-tuned with only the target task-specific dataset. For example, when the ULMFit was fine-tuned with Pubmed+Cadec+AskAPatient, the average accuracy tested on AskAPatient is 78.17% compared to 67.25% when the model was fine-tuned with Pubmed+Cadec, as shown in Table VIII, and it is also higher than 77.16% when the model was fine-tuned with only AskAPatient, as shown in Table VII. The same pattern also exists for dataset TwADR-L, as shown in Tables VII and IX.
4) BERT achieved the best normalization accuracy on both datasets AskAPatient and TwADR-L. Its average accuracy of the normalization on dataset AskApatient is 84.91%. Compared to 78.17%, the highest accuracy of ULMFit, it is about 8.6% increase, as shown in Table VIII. Its average accuracy of the normalization on dataset TwADR-L is 41.71%. Compared to 39.85%, the highest accuracy of ULMFit, it is about 4.7% increase, as shown in Table IX.

The results of the experiments provide us concrete evidences that fine-tuning could improve the performance of machine learning. For example, when the combination dataset Pubmed+Cadec+AskAPatient was used to fine-tune the language model of ULMFit, the average accuracy of the normalization on dataset AskAPatient was increased to 78.17% from 65.07% when the model was not fine-tuned. However, the improvement is not guaranteed, and it may produce a negative impact on the performance, as shown by the results of experiments in this section. For example, when the language model of ULMFit was fine-tuned with dataset Big_tweet, the average accuracy of normalization on dataset AskAPatient was decreased from the original 65.07% to only 22.65%. Therefore, carefully choosing the dataset for the fine-tuning is very important to performance improvement. For example, when dataset Pubmed+Cadec was used for fine-tuning the language model, the average accuracy is higher than any other experiment that does not include the dataset. However, the dataset Big_tweet alone always resulted to the lowest accuracy. When a combination dataset includes the target task-specific dataset, the average accuracy for each configuration is very close to or better than the result of experiment that was fine-tuned only with the target task-specific dataset. The same pattern is consistent in both datasets AskAPatient and TwADR-L. The results of the experiments clearly demonstrated the impact of the quality of the dataset to the performance of the fine-tuning.

Based on the aforementioned discussion and the review of these datasets, we summarize the observation: If a machine learning model is pretrained with a dataset (called source dataset) for general tasks, such as text classification, and then it is fine-tuned with another dataset (called target dataset) for a specific task, such as medical concept normalization, then we expect that the target specific task is an instance of the source general tasks. Therefore, we also expect that the fine-tuning process would adapt the source dataset into an extended target dataset so the source dataset and the target dataset become the subsets of the extended target dataset. The fine-tuning would increase the quality of the target dataset by taking advantage of the information brought by the source dataset.

However, if a dataset that is chosen for fine-tuning the model is not related to the target-specific task, then the fine-tuning may produce negative performance impact since the fine-tuning

process would consider the one as a target-specific dataset and give it more attention/weight during training. For example, when ULMFit model was fine-tuned only with Pubmed, the model was tuned with disproportionately more formal medical concepts than normal cases, which might mislead the training and increase the bias of the trained model. The same explanation can be applied to dataset Big_tweet, which distorted the source dataset with manually added disproportionately informal medical phrases. Therefore, when fine-tuning is applied to a pretrained model, it is necessary to conduct an evaluation of the target datasets to ensure they are appropriate to the target-specific task.

BERT produced the best result due to its powerful model and the quality of its training data. Its training data cover much more formal and informal medical concepts than the training dataset used for training the language model AWD-LSTM of ULMFit. We believe the general text for training BERT may contain most of the phrases and medical concepts in datasets Pubmed, Healthnews, Cadec, and Big_tweet. Therefore, it has fewer number of concepts or phrases that have used the default embedding vectors, which cannot accurately capture the semantics of the concepts or phrases. In addition, BERT uses a deep bidirectional contextual representation so that each word is contextualized using both words to its left and its right. Therefore, word embeddings produced by BERT could much better represent the relation among words, which is critical for medical concept normalization.

V. Data Quality Evaluation for Machine Learning

Since the data quality can impact the performance of a machine learning system that is built on the data, it is necessary to evaluate the data quality systematically. There are many approaches for data quality evaluation [37], [38], but few work directly related to building machine learning systems. In this section, we discuss how to systematically evaluate the data quality to ensure the quality of machine learning, specifically of deep learning. Based on discussions in previous sections and other data quality research results, especially paper [17] authored by Wang and Strong, we propose three quality attributes that are most critical to the "fit for purposes" of deep learning. The three quality attributes are: *comprehensiveness*, *correctness*, and *variety*, the three most important quality attributes to the performance of machine learning [39]. The definitions could be slightly different from the definitions in other data quality publications, including Wang and Stron [17] since our focus is on the "fit for purpose" of building a machine learning system.

A. Data Quality Attributes

Many examples have demonstrated the importance of the correctness of training data for a machine learning system. One of the examples is to use generative adversarial network (GAN) to produce label noises into training data to easily degrade the performance of a machine learning system [40]. It is obvious that label noises existing in a validation dataset could produce an overfitted model, and it could produce a misleading testing result when they exist in a test dataset. When a machine learning model is trained with a dataset that suffers the comprehensiveness issue, then the trained model could suffer the generalization issue. Therefore, the state-of-the-art language models, such as BERT or GPT-3 [41], were trained with billions of words. Variety emphasizes the uniqueness of each type of dataset, which is important to ensure the confidence of the evaluation results. It is well understood that the large overlap existing between training data and test data could produce inflation results.

Although many quality attributes were proposed for data quality evaluation. For example, quality attributes, such as completeness, timeliness, consistency, and volume, are also important for measuring the quality of data. However, our discussion is limited to the only quality attributes as they are important to the quality of machine learning. The other quality attributes either are not important to the "fit for purpose" of building a machine learning system or can be considered as the properties of the three quality attributes. For example, completeness can be substituted by comprehensiveness, timeliness, and volume are related to comprehensiveness and variety, and consistency is part of correctness.

"Comprehensiveness" means a dataset contains all representative samples from the population. For example, the machine learning project for the medical concept normalization in ADR should include all medical concepts and their corresponding informal phrases related to ADR. The importance of the comprehensiveness of data to machine learning, especially deep learning, is well understood since a deep learning model normally includes millions of parameters that needs a large amount of data to train it. For example, the ImageNet includes 14 million images in 22 000 categories, and it has been widely used for training computer vision-related deep learning models. BERT was pretrained with more than three billion words, and it is continually retrained with more words. A deep learning model is trained with a more comprehensive training dataset may achieve better performance than a model that is trained with a less comprehensive model. The problem that produced a disparate performance for different ethnicity groups reported in [6] was solved when a more comprehensive dataset was used for training the machine learning model. BERT was trained with a much more comprehensive dataset comparing to the dataset used for training AWD-LSTM, and it might be one of the reasons the BERT-based medical concept normalization achieved better performance.

"Correctness" refers to the fact that a record in a dataset is accurate and valid, and they are correctly labeled if they are labeled records. Inaccurate or invalid data lead to data noises, and incorrectly labeled data lead to label noises. Therefore, a correct dataset should contain minimal label noises and data noises. For example, the medical concepts should be collected from reliable sources, and every mapping between a Twitter phrase and a medical concept in a dataset should be correct. If a dataset includes too many incorrect data items, then the model built on the dataset could be incorrectly trained or evaluated. For example, the problem of ChexNet [7] we discussed in Section I was due to the incorrect assumption of the labels of unlabeled images [8].

"Variety" is about the coverage of a dataset of all different cases on selected features. In this sense, the variety is a subset of comprehensiveness, and it is a quality attribute for sampling. The variety requires the distribution of a dataset in a feature to be as similar as the distribution of its population. For example, if a dataset is supposed to be normal distribution in a selected feature with known mean and standard deviation, then the real dataset should be normal distribution with similar mean value and standard deviation in the same feature. The distribution of a population could be known in advance, or they can be approximated through simulation, sampling, or bootstrapping. If more than 99% of data items in a dataset are collected within one standard deviation of a population that is normal distribution in a feature, then the dataset has low variety since it represents less than 68% of the population. In addition, the variety requires each validation dataset and test dataset contains a significant amount of new data comparing to the corresponding training dataset. The percentage of the overlapped data between a test/validation dataset and its corresponding training dataset should be as low as possible, such as less than 10%.

We now discuss the approach for evaluating the comprehensiveness, correctness, and variety of datasets using the medical concepts normalization project as an example.

B. Check the Comprehensiveness of a Dataset

The quality attribute comprehensiveness of a dataset is correlated to the quality of medical concept normalization. If the dataset for training the language model does not contain a sufficient number of medical concepts or informal phrases that are related to medical concepts, some of the concepts and phrases to be tested in the normalization will not be embedded with accurate word vectors that capture their real meanings. Those concepts and phrases are represented with default vectors that are generated by the general language model. For example, the "cell" in "biology cell" could be embedded as the same word vector as the "cell" in "cell phone." If many medical concepts and phrases use the default word vectors, the normalization built on the word embedding is unlikely to achieve high accuracy, as the normalization basically is built on the measurement of the distance of the vectors of the informal phrase and its corresponding medical concept. Each word used in social media may have many variations and typos, and a large number of words could be used in social media to describe a specific topic, such as ADR discussions. Therefore, it is crucial to understand the population so that we can measure the comprehensiveness of a dataset to be used for building the machine learning system. Domain knowledge is essential to understand the comprehensiveness of the population. For the medical concept normalization example, we first know the population includes all ADR-related concepts, and related informal phrases are appearing in top ADR social media and online forums over a period of time.

In order to build a comprehensive dataset for the medical concept normalization, one may need first to collect initial related Twitter messages (if we only consider Twitter messages) and extract unique phrases from the messages. Then, randomly collect more related Twitter messages and extract unique phrases from new messages or historical messages in different sites. The new phrases collected are compared to the phrase collection built before. If a new phrase is found, it is added to the phrase collection and the process continues until a few new phrases will be found. Building comprehensive medical concepts can be done with a domain expert to collect concepts from medical documents, dictionaries, and medical ontology. Data augmentation could also be used to produce new phrases and medical concepts.

Dataset AskAPatient only includes 3749 unique phrases and 1036 medical concepts. Dataset TwADR-L contains only 1436 unique phrases and 2220 medical concepts, but 1947 of them do not map to any phrase. According to the side effective resource database SIDER 4.1,[4] there are 5868 unique side effects related to ADR. Each side effect is defined by at least one medical concept. Therefore, the comprehensiveness of the two datasets is fairly low. The results of the experiments presented in the last section demonstrated that transfer learning is an effective approach for improving the comprehensiveness of a dataset. It "transfers" a much more comprehensive dataset that was used for pretraining the language model into a small domain-specific dataset to improve its comprehensiveness [4], [42].

Domain knowledge is required to evaluate the comprehensiveness of a dataset. One way of the evaluation is to evaluate the data collection procedure and data sources. For example, we may look at ImageNet as a comprehensive dataset for building a regular image classification system. If a dataset for medical concept normalization was produced following the procedure we just talked about, it could be looked at as a comprehensive dataset. When we evaluate the comprehensiveness of a dataset for machine learning, it would be sufficient to measure the comprehensiveness by quality rather than its quantity.

C. Check the Correctness of a Dataset

The validity and accuracy of data usually can be checked using exploratory data analysis tools. For example, a Q–Q plot could identify the outliers and skewness in a dataset. Checking the label noise could be complicated, especially for a large-scale dataset, such as AskAPatient and ImageNet. The most straightforward way is to check the sample data manually. However, machine learning algorithms could be used for checking label noises. For example, machine learning methods, such as K-means clustering and Gaussian mixture model, can be used to group the labeled data, and then one can check the labeling of outliers and data items located in the overlapped clustering areas for any label noises [39].

D. Check the Variety of a Dataset

This task is to check whether a dataset has a distribution in selected parameters as the expected distribution of the population, and the new data in training, validation, and test datasets. For example, dataset AskAPatient contains many duplications, and the number of duplications of a phrase should reflect the frequency of that phrase appearing in the collected text, such

[4][Online]. Available: http://sideeffects.embl.de/

as Twitter messages and medical forums. We expect that the duplication pattern of the phrases in the dataset is consistent with the usage frequency pattern in the data sources. Therefore, simply removing the duplicated data items in the dataset might not be an appropriate way of cleaning the dataset [19]. Instead, one can calculate the usage frequency of a phrase in the collected data and then compare it with the duplication of the corresponding record in the dataset. If we build a histogram of selected phrases and then calculate its distribution of a dataset, we can compare the distribution of the dataset to the distribution of the population in the same phrases. The distribution of a population is either known in advance or able to be approximated using simulation or bootstrapping. Variety checking would also provide us the sample size and the distribution information, such as mean and standard deviation of the dataset, which is essential to the statistical inference of the machine learning result.

The other property of variety that needs to be checked is the unique data items in a dataset. Some statistical tools, such as IBM SPSS and GNU PSPP, can be used to check the uniqueness in a dataset based on statistical models, such as linear regression and visualizations, such as Q–Q plot [37]. Checking variety is also essential to the medical concept normalization. For example, it is necessary to know how many unique medical concepts and informal phrases exist in a dataset, as well as labeled records. If the unique medical concepts are significantly less than expected in a dataset, or lots of the informal phrases are not labeled, or many medical concepts are not mapped to a phrase, then the quality of the variety of the dataset is low, it may produce a poor quality system.

The third property of variety needs to be checked is the overlap in training, validation, and test datasets. Checking the overlap among datasets is straightforward through comparing the data items among the datasets. However, it could be complicated for complex data, such as videos, audios, and images. In those cases, the comparison could be conducted on the identities of the data items or convert the data item into a value (such as a hash value) that can be compared. As the experiment results are shown in this article, overlap in the test dataset could cause inflation of the system performance evaluation. Therefore, it is necessary to calculate the overlap of datasets when one validates a machine learning system. Although the correctness of a dataset is important for building a machine learning system, adversarial data are also important for testing the reliability of a system. Adversarial data, such as label noise that are generated on purpose using techniques, such as GAN, might be needed to be included in test data. We consider the percentage of adversarial data in a test dataset as one measurement of the variety.

As the results of experiments shown in the last section, transfer learning is an effective way of improving the variety of a dataset provided the source dataset and the target dataset are appropriately selected.

VI. Related Work

The methods used in medical concept normalization can be divided into three categories: string-matching methods, rule-based approaches, and deep learning algorithms.

String-matching methods identify concepts according to multiple resources, whereas rule-based approaches map medical concepts based on rules. Li *et al.* [43] proposed a rule-based approach that generated candidates for a given biomedical entity using three types of rules. Kang *et al.* [44] proposed to use rule-based natural language processing to improve normalization performance, where five rules were applied to address specific tasks, such as coordination, abbreviation, and term variation. Some researchers proposed a solution that combines several approaches together. For example, Dogan and Lu [45] combined a string-matching method and a rule-based approach to cross-validate the identified results.

Recently, semantic information and deep learning methods have been proven effective in medical concept normalization [43], especially among social media text. For example, in order to map from social media message "I am not calm or easy" to the medical concept "Agitate," normalization has to take into account the semantics of the whole message; otherwise, the text may be mapped to the medical concept "calm." Normalization systems have been developed to learn and exploit the semantic similarity between text from social media messages and medical concepts using deep neural networks, such as CNN and RNN.

Leaman *et al.* introduced machine learning to the medical concept normalization task recently [46]. The method can learn similarities between phrases and concept names directly from training data, which proved to be effective and has been served as a baseline to other medical concept normalization studies. Belousov *et al.* [47] proposed an ensemble system that combines generalized linear and deep learning models trained on both generic and target domain word embeddings in SMM4H 2017 medical concept normalization task, and the system achieved high accuracy on the test dataset. Luo *et al.* [48] argued that traditional CNN could hardly capture matching signals. They developed a multiview learning, which included one CNN for each view, and then the outputs from the CNNs are combined. Their experiment was conducted on a disease dataset from a Chinese hospital and achieved impressive accuracy.

Lee *et al.* [19] conducted medical concept normalization for online user-generated text based on TwADR-L and AskAPatient datasets. However, the overlapped data among test and training data could contribute to the false improvement of the normalization accuracy. Lee *et al.* [19] cleaned and recreated the training, validation, and test datasets and removed all medical concepts that had less than five examples. The results of their experiments showed that CNN achieved 19.46% and 55.46% on accuracy on TwADR-L and AskAPatient, respectively, RNN achieved 25.30% and 65.04% on accuracy on TwADR-L and AskAPatient, respectively. In addition, word embedding trained on health-related Tweets messages had the most significant impact on the classification performance [19], which showed that fine-tuning the word embedding is a potential approach for improving the quality of datasets.

Niu *et al.* [49] proposed a multitask character-level attention network to normalize standard medical concepts in social media messages. The character-level encoding scheme can capture character-level features even in out-of-vocabulary words, whereas the word-level morphological information in the

medical concept is effectively exploited to supervise the training of an auxiliary network. They also conducted experiments on TwADR-L and AskAPatient datasets. The results of the experiments showed the proposed multitask attentional character-level CNN achieved impressive performance on TwADR-L and AskAPatient datasets, which were 46.46% and 84.65% on accuracy, respectively [49].

Another related work of our research is the evaluation of data quality. Data validation is an essential requirement to ensure the quality of machine learning systems, and low-quality data may cause problems, such as a wrong prediction or low classification accuracy [50]. Datasets that were not adequately evaluated can produce misleading results. Gao *et al.* discussed big data quality issues, challenges, and evaluated tools for validation and quality assurance of big data [51]. Wang and Strong [17] developed a hierarchical framework for defining data quality attributes in four categories: intrinsic data quality, contextual data quality, representational data quality, and accessibility data quality. They concluded that "high-quality data should be intrinsically good, contextually appropriate to the task, clearly represented, and accessible to the data consumer" [17]. The framework is still applied to the data for machine learning in general. Chen *et al.* proposed a comprehensive and practical framework to evaluate data quality, especially complex data, such as knowledge graphs [15]. The focus of our work is on the contextual data quality of datasets for deep learning. Through a case study of medical concept normalization, we demonstrated the impact of the quality of datasets on the performance of deep learning. We defined three quality attributes regarding the data comprehensiveness, variety, and correctness that would impact the performance of deep learning.

Although many publications on data quality have been published, the discussion on the impact of data quality on the machine learning performance with evidence and measurement is rare. We identified three data quality attributes that are most important to deep learning—comprehensiveness, correctness, and variety. Batini *et al.* [37] gave a comprehensive review of the methodologies for data quality assessment and improvement, but the application domain is limited to database applications. Crowdsourcing is an effective and widely adopted way of collecting a large amount of data for deep learning [3]. The results reported in the paper [3] showed data with noises could still be effective for training a deep learning model. But we believe if noises are removed from a training dataset, the performance of the deep learning trained with the dataset could be even higher. Sun *et al.* [12] have shown the impact of the size and quality of a dataset on the performance of deep learning.

Some preliminary work on the data quality evaluation on domain-specific applications was reported recently [15], [39]. Evaluation of the correctness of web sources using hyperlinks, browsing history, and the factual information provided by the source was reported [52]. Some evaluations were conducted based on the relationship between web sources and their information [53]. Finding duplicates in a dataset is also an important quality assurance task in machine learning. Machine learning algorithms, such as gradient boosted decision tree, have been used for detecting duplicates [54]. Data filtering is an approach for quality assurance by removing bad data from data sources. Nobles *et al.* conducted an evaluation of the completeness and availability of electronic health record data. They identified undesirable data in datasets using machine learning algorithms, such as support vector machine and deep learning. Since the label noise could reduce the performance of machine learning, one needs to either improve the machine learning algorithm to handle the noise or improve the quality of the data through filtering the noise [55]. Foidl and Felderer proposed three criteria, including data source quality, data smells, data pipeline quality to identify the low-quality data. However, they only presented a conceptual approach without validating it using a real-world case.

Transfer learning has been widely used for improving the performance of machine learning across many tasks, and transfer learning by fine-tuning pretrained neural networks outperforms the networks that are trained from scratch on the same data [56]. It is almost a standard procedure to train a deep learning model through fine-tuning a model that has been pretrained with a large-scale dataset, such as ImageNet. For example, pretrained AlexNet was fine-tuned for the classification of biomedical images [57]–[59]. To compare the performance of fine-tuning, CIFAR-10 and CIFAR-100 were used as source and target datasets for the transfer learning in [60]. Recently, Lu *et al.* proposed a neural architecture transfer, a fine-tuning-based transfer learning for image classification that leveraged an existing supernet and efficiently transferred it into a task-specific supernet, showed great scalability and practicality in different scenarios [61]. In the text domain, ULMFit and BERT are the two most frequently models that use the idea of fine-tuning. The ULMFit model consists of three stages: language model pretrained on a general-domain corpus, language model fine-tuned using domain-specific unlabeled data, and classifier fine-tuned on the target task using gradual unfreezing technique [28]. BERT is another pretrained model that can be fine-tuned with an input to produce context-aware word embedding [29]. In order to show the performance of different transfer learning models, Zhuang *et al.* conducted an experimental study on 20 representative transfer learning models, and its results demonstrated that appropriate transfer learning models should be carefully selected for different applications [62].

VII. Conclusion

Both the quality of a dataset and the capability of a model could contribute on the performance of a machine learning system. However, the research for understanding the impact of the quality of datasets to the performance of machine learning is just emerging. In this article, we first introduced an experimental study to illustrate how noises in datasets could contribute to the false performance improvement that could exist in many machine learning systems. Then, we experimented with a transfer learning approach for improving the quality of datasets and demonstrated the true performance improvement of the machine learning system that was trained and tested on the datasets. However, transfer learning is not always effective for quality improvement. Therefore, a group of guidelines were proposed for using transfer learning, which are as follows.

1) The datasets for fine-tuning should be related to the target-specific task, otherwise the fine-tuning may produce negative performance impact.
2) The fine-tuning can be used to improve data quality by taking advantage of the information brought by the source dataset.
3) A powerful model, such as BERT, can better capture semantic information, thereby can be the prior model for fine-tuning.

In order to evaluate the quality of datasets, we proposed three data quality criteria and the approaches for measuring them. We explained the research problem and results through studying a machine learning system for normalizing medical concepts in social media text with widely adopted open datasets.

In the future, we will conduct experiments on transfer learning of complex data, such as knowledge graphs, to investigate whether a domain specific knowledge graph can be effectively transferred with general knowledge graph, such as WikiData, and define test adequacy criteria on the quality of datasets for testing machine learning systems.

ACKNOWLEDGMENT

The authors would like to thank Dr. S. Wang at Tianjin Normal University and J. Li at the University of North Texas for the assistance of the experimental study. The authors are grateful to all the anonymous reviewers for their precious comments and suggestions.

REFERENCES

[1] Y. Bengio, "Learning deep architectures for AI," *Found. Trends Mach. Learn.*, vol. 2, no. 1, pp. 1–127, 2009.
[2] D. Silver *et al.* "A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play," *Science*, vol. 362, no. 6419, pp. 1140–1144, 2018.
[3] J. Krause *et al.*, "The unreasonable effectiveness of noisy data for fine-grained recognition," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 301–320.
[4] R. Raina, A. Battle, H. Lee, B. Packer, and A. Y. Ng, "Self-taught learning: Transfer learning from unlabeled data," in *Proc. 24th Int. Conf. Mach. Learn.*, 2007, pp. 759–766.
[5] D. Rolnick, A. Veit, S. Belongie, and N. Shavit, "Deep learning is robust to massive label noise," *CoRR*, vol. abs/1705.10694, 2017: Accessed: 19 Apr. 2021. [Online]. Available: http://arxiv.org/abs/1705.10694
[6] J. Buolamwini and T. Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification," in *Proc. Conf. Fairness, Accountability, Transparency*, 2018, pp. 77–91.
[7] P. Rajpurkar *et al.*, "CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning," *CoRR*, vol. abs/1711.05225, 2017. Accessed: 19 Apr. 2021. [Online]. Available: http://arxiv.org/abs/1711.05225
[8] L. Oakden-Rayner, "Exploring the ChestXray14 dataset: Problems," 2017. Accessed: Jun. 19, 2019. [Online]. Available: https://lukeoakdenrayner.wordpress.com/2017/12/18/the-chestxray14-dataset-problems
[9] A. Eklund, T. E. Nichols, and H. Knutsson, "Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates," *Proc. Nat. Acad. Sci. USA*, vol. 113, no. 28, pp. 7900–7905, 2016.
[10] E. Giannoulatou, S.-H. Park, D. Humphreys, and J. Ho, "Verification and validation of bioinformatics software without a gold standard: A case study of BWA and Bowtie," *BMC Bioinformat.*, vol. 15(*Suppl*16), 2014, Art. no. S15.
[11] J. Zhang *et al.*, "Analysis of cellular objects through diffraction images acquired by flow cytometry," *Opt. Exp.*, vol. 21, no. 21, pp. 24819–24828, 2013.
[12] C. Sun, A. Shrivastava, S. Singh, and A. Gupta, "Revisiting unreasonable effectiveness of data in deep learning era," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 843–852.
[13] C. Su and D. Huang, "Hybrid recommender system based on deep learning model," *Int. J. Performability Eng.*, vol. 16, no. 1, pp. 118–129, 2020.
[14] D. Firmani, M. Mecella, M. Scannapieco, and C. Batini, "On the meaningfulness of 'big data quality'," *Data Sci. Eng.*, vol. 1, no. 1, pp. 6–20, 2016.
[15] H. Chen, G. Cao, J. Chen, and J. Ding, "A practical framework for evaluating the quality of knowledge graph," in *Proc. China Conf. Knowl. Graph Semantic Comput.*, 2019, pp. 111–122.
[16] J. R. Evans and W. M. Lindsay, *The Management and Control of Quality*, vol. 5. Cincinnati, OH, USA: South-Western, 2002, pp. 115–128.
[17] R. Y. Wang and D. M. Strong, "Beyond accuracy: What data quality means to data consumers," *J. Manage. Inf. Syst.*, vol. 12, no. 4, pp. 5–33, 1996.
[18] N. Limsopatham and N. Collier, "Normalising medical concepts in social media texts by learning semantic representation," in *Proc. 54th Annu. Meeting Assoc. Comput. Linguistics (Vol. 1: Long Papers)*, 2016, pp. 1014–1023.
[19] K. Lee, S. A. Hasan, O. Farri, A. Choudhary, and A. Agrawal, "Medical concept normalization for online user-generated texts," in *Proc. IEEE Int. Conf. Healthcare Informat.*, 2017, pp. 462–469.
[20] C. Tan, F. Sun, T. Kong, W. Zhang, C. Yang, and C. Liu, "A survey on deep transfer learning," in *Proc. Int. Conf. Artif. Neural Netw.*, 2018, pp. 270–279.
[21] ILSVRC2012, "ImageNet large scale visual recognition challenge," 2012. Accessed: Jun. 20, 2019. [Online]. Available: http://www.image-net.org/challenges/LSVRC/2012/results.html
[22] Y. Wand and R. Y. Wang, "Anchoring data quality dimensions in ontological foundations," *Commun. ACM*, vol. 39, no. 11, pp. 86–95, 1996.
[23] O. Bodenreider, "The Unified Medical Language System (UMLS): Integrating biomedical terminology," *Nucleic Acids Res.*, vol. 32, no. suppl_1, pp. D 267–D270, 2004.
[24] D. Britz, "Implementing a CNN for text classification in TensorFlow," Accessed: Jan. 10, 2020. [Online]. Available: http://www.wildml.com/2015/12/implementing-a-cnn-for-text-classification-in-tensorflow/
[25] "Recurrent neural networks and LSTM tutorial in Python and TensorFlow." 2017. Accessed: Jan. 10, 2020. [Online]. Available: https://adventuresinmachinelearning.com/recurrent-neural-networks-lstm-tutorial-tensorflow/
[26] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 3111–3119.
[27] E. C. Too, L. Yujian, S. Njuki, and L. Yingchun, "A comparative study of fine-tuning deep learning models for plant disease identification," *Comput. Electron. Agriculture*, vol. 161, pp. 272–279, 2019.
[28] J. Howard and S. Ruder, "Universal language model fine-tuning for text classification," *CoRR*, vol. abs/1801.06146, 2018. Accessed: 19 Apr. 2021. [Online]. Available: http://arxiv.org/abs/1801.06146
[29] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-trainingof deep bidirectional transformers for language understanding," *CoRR*, vol. abs/1810.04805, 2018. Accessed 19 Apr. 2021. [Online]. Available: http://arxiv.org/abs/1810.04805
[30] "The TwADR-L dataset." 2016. Accessed: Feb. 10, 2020. [Online]. Available: https://zenodo.org/record/55013#.XK-1zOtKh24
[31] "Healthcare Twitter analysis." 2016. Accessed: Feb. 10, 2020. [Online]. Available: https://github.com/grfiv/healthcare_twitter_analysis
[32] C. Sun, X. Qiu, Y. Xu, and X. Huang, "How to fine-tune bert for text classification?" in *Proc. China Nat. Conf. Chin. Comput. Linguistics*, 2019, pp. 194–206.
[33] "The WikiText long term dependency language modeling dataset—WikiText-103." 2020. Accessed: Jan. 10, 2020. [Online]. Available: https://blog.einstein.ai/the-wikitext-long-term-dependency-language-modeling-dataset/
[34] J. Pennington, R. Socher, and C. D. Manning, "GloVe: Global vectors for word representation," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2014, pp. 1532–1543.
[35] S. Faltl, M. Schimpke, and C. Hackober, "ULMFiT: State-of-the-art in text analysis." 2019. Accessed: Jan. 10, 2020. [Online]. Available: https://humboldt-wi.github.io/blog/research/information_systems_1819/group4_ulmfit/
[36] S. Merity, N. S. Keskar, and R. Socher, "Regularizing and optimizing LSTM language models," *CoRR*, vol. abs/1708.02182, 2017. Accessed: 19 Apr. 2021. [Online]. Available: http://arxiv.org/abs/1708.02182
[37] C. Batini, C. Cappiello, C. Francalanci, and A. Maurino, "Methodologies for data quality assessment and improvement," *ACM Comput. Surv.*, vol. 41, no. 3, pp. 1–52, 2009.

[38] F. Daniel, P. Kucherbaev, C. Cappiello, B. Benatallah, and M. Allahbakhsh, "Quality control in crowdsourcing: A survey of quality attributes, assessment techniques, and assurance actions," *ACM Comput. Surv.*, vol. 51, no. 1, pp. 1–40, 2018.

[39] J. Ding, X. Li, X. Kang, and V. N. Gudivada, "A case study of the augmentation and evaluation of training data for deep learning," *J. Data Inf. Qual.*, vol. 11, no. 4, pp. 1–22, 2019.

[40] T. Bai *et al.*, "AI-GAN: Attack-inspired generation of adversarial examples," *CoRR*, vol. abs/2002.02196, 2020. Accessed: 19 Apr. 2021. [Online]. Available: https://arxiv.org/abs/2002.02196

[41] T. B. Brown *et al.*, "Language models are few-shot learners," 2020, *arXiv:2005.14165*.

[42] L. Torrey and J. Shavlik, "Transfer learning," in *Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques*. Hershey, PA, USA: IGI Global, 2010, pp. 242–264.

[43] H. Li *et al.*, "CNN-based ranking for biomedical entity normalization," *BMC Bioinformat.*, vol. 18, no. 11, pp. 79–86, 2017.

[44] N. Kang, B. Singh, Z. Afzal, E. M. van Mulligen, and J. A. Kors, "Using rule-based natural language processing to improve disease normalization in biomedical text," *J. Amer. Med. Informat. Assoc.*, vol. 20, no. 5, pp. 876–881, 2013.

[45] R. I. Dogan and Z. Lu, "An inference method for disease name normalization," in *Proc. AAAI Fall Symp. Ser.*, 2012, pp. 8–13.

[46] R. Leaman, R. Islamaj Doğan, and Z. Lu, "DNorm: Disease name normalization with pairwise learning to rank," *Bioinformatics*, vol. 29, no. 22, pp. 2909–2917, 2013.

[47] M. Belousov, W. Dixon, and G. Nenadic, "Using an ensemble of generalised linear and deep learning models in the SMM4H 2017 medical concept normalisation task," in *Proc. 2nd Workshop Social Media Mining Health Appl.*, 2017, pp. 54–58.

[48] Y. Luo, G. Song, P. Li, and Z. Qi, "Multi-task medical concept normalization using multi-view convolutional neural network," in *Proc. AAAI Conf. Artif. Intell.*, 2018, pp. 5868–5875.

[49] J. Niu, Y. Yang, S. Zhang, Z. Sun, and W. Zhang, "Multi-task character-level attentional networks for medical concept normalization," *Neural Process. Lett.*, vol. 49, no. 3, pp. 1239–1256, 2019.

[50] H. Foidl and M. Felderer, "Risk-based data validation in machine learning-based software systems," in *Proc. 3rd ACM SIGSOFT Int. Workshop Mach. Learn. Techn. Softw. Qual. Eval.*, 2019, pp. 13–18.

[51] J. Gao, C. Xie, and C. Tao, "Big data validation and quality assurance—Issuses, challenges, and needs," in *Proc. IEEE Symp. Serv.-Oriented Syst. Eng.*, Mar. 2016, pp. 433–441.

[52] X. L. Dong *et al.*, "Knowledge-based trust: Estimating the trustworthiness of web sources," *CoRR*, vol. abs/1502.03519, 2015. Accessed: 19 Apr. 2021. [Online]. Available: http://arxiv.org/abs/1502.03519

[53] X. Yin, J. Han, and S. Y. Philip, "Truth discovery with multiple conflicting information providers on the web," *IEEE Trans. Knowl. Data Eng.*, vol. 20, no. 6, pp. 796–808, Jun. 2008.

[54] C. H. Wu and Y. Song, "Robust and distributed web-scale near-dup document conflation in Microsoft academic service," in *Proc. IEEE Int. Conf. Big Data*, Oct. 2015, pp. 2606–2611.

[55] J. A. Sáez, B. Krawczyk, and M. Woźniak, "On the influence of class noise in medical data classification: Treatment using noise filtering methods," *Appl. Artif. Intell.*, vol. 30, no. 6, pp. 590–609, 2016.

[56] S. Kornblith, J. Shlens, and Q. V. Le, "Do better ImageNet models transfer better?" in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 2661–2671.

[57] H.-C. Shin *et al.*, "Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning," *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1285–1298, May 2016.

[58] W. Nawaz, S. Ahmed, A. Tahir, and H. A. Khan, "Classification of breast cancer histology images using ALEXNET," in *Proc. Int. Conf. Image Anal. Recognit.*, 2018, pp. 869–876.

[59] Z. Zhou, J. Shin, L. Zhang, S. Gurudu, M. Gotway, and J. Liang, "Fine-tuning convolutional neural networks for biomedical image analysis: Actively and incrementally," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 7340–7351.

[60] A. Shahriari, "Distributed deep transfer learning by basic probability assignment," *CoRR*, vol. abs/1710.07437, 2017. Accessed: 19 Apr. 2021. [Online]. Available: http://arxiv.org/abs/1710.07437

[61] Z. Lu, G. Sreekumar, E. Goodman, W. Banzhaf, K. Deb, and V. N. Boddeti, "Neural architecture transfer," *IEEE Trans. Pattern Analysis Mach. Intell.*, early access, 2021, pp. 1–1.

[62] F. Zhuang *et al.*, "A comprehensive survey on transfer learning," in *Proc. IEEE*, vol. 109, no. 1, 2020, pp. 43–76.