When BERT Meets Bilbo:
A Learning Curve Analysis of Pretrained Language Model on Disease Classification

Xuedong Li*, Walter Yuan[†], Dezhong Peng*, Qiaozhu Mei[‡], Yue Wang[§]

*College of Computer Science, Sichuan University, Chengdu, China
[†]MobLab Inc., Pasadena, CA
[‡]School of Information, University of Michigan, Ann Arbor, MI
[§]School of Information and Library Science, University of North Carolina at Chapel Hill
Email: *xuedong.li@stu.scu.edu.cn, [†]walter.yuan@moblab.com, *pengdz@scu.edu.cn,
[‡]qmei@umich.edu, [§]wangyue@email.unc.edu

Abstract—Objective: Natural language processing (NLP) tasks in the health domain often deal with limited amount of labeled data due to high annotation costs and naturally rare observations. To compensate for the lack of training data, health NLP researchers often have to leverage knowledge and resources external to a task at hand. Recently, pretrained large-scale language models such as the Bidirectional Encoder Representations from Transformers (BERT) have been proven to be a powerful way of learning rich linguistic knowledge from massive unlabeled text and transferring that knowledge to downstream tasks. However, previous downstream tasks often used training data at such a large scale that is unlikely to obtain in the health domain. In this work, we aim to study whether BERT can still benefit downstream tasks when training data are relatively small in the context of health NLP.[1]

Method: We conducted a learning curve analysis to study the behavior of BERT and baseline models as training data size increases. We observed the classification performance of these models on two disease diagnosis data sets, where some diseases are naturally rare and have very limited observations (fewer than 2 out of 10,000). The baselines included commonly used text classification models such as sparse and dense bag-of-words models, long short-term memory networks, and their variants that leveraged external knowledge. To obtain learning curves, we incremented the amount of training examples per disease from small to large, and measured the classification performance in macro-averaged F_1 score.

Results: On the task of classifying all diseases, the learning curves of BERT were consistently above all baselines, significantly outperforming them across the spectrum of training data sizes. But under extreme situations where only one or two training documents per disease were available, BERT was outperformed by linear classifiers with carefully engineered bag-of-words features.

Conclusion: As long as the amount of training documents is not extremely few, fine-tuning a pretrained BERT model is a highly effective approach to health NLP tasks like disease classification. However, in extreme cases where each class has only one or two training documents and no more will be available, simple linear models using bag-of-words features shall be considered.

Index Terms—learning curve; bidirectional encoder representations from transformers; disease classification

---

[1]In *The Lord of the Rings*, "Bert" is a giant stone troll. In terms of size, health NLP data are the "Hobbits" among machine learning data sets.

I. INTRODUCTION

Machine learning has become the predominant approach to health natural language processing (NLP) in recent years. To achieve high performance, machine learning models often need to be trained on a substantial amount of labeled data. Deep learning models, while capable of achieving even higher performance, may need more training data to train a large number of internal parameters.

Unlike NLP tasks in the general domain where training data are abundant or can be collected at scale, preparing training data in the health domain can be prohibitively expensive and even infeasible. For instance, labeling social media posts can be crowdsourced at a very low cost through Amazon Web Services [1], while annotating clinical notes requires special medical training and long hours [2]. On the task of rare disease identification, the amount of labeled documents is further bounded by the size of population, since rare diseases appear very infrequently (a rare disease affects fewer than 1 in 1,500 people in the U.S. [3], or 1 in 2,000 in Europe [4]). As a result, health NLP researchers have been proposing a variety of methods to compensate for the lack of training data [5]. These include leveraging expert knowledge and medical ontologies [6]–[8], transferring statistical knowledge learned from related tasks [9], simultaneously learning from multiple tasks [10], and selectively asking experts for labels [11].

Recently, Bidirectional Encoder Representations from Transformers (BERT) model has been increasingly adopted by the NLP research community as it celebrates superior performance in a wide range of NLP tasks [12]. BERT learns contextual representation of words using information from both sides of a word, effectively capturing syntactic and semantic knowledge that can benefit many NLP tasks. A pretrained BERT model can be tailored to a specific NLP task by using the task-specific data to further train the model, a procedure known as "fine-tuning". In this way, the new task can build on top of the pretrained knowledge in BERT to achieve superior generalization performance. However, previ-
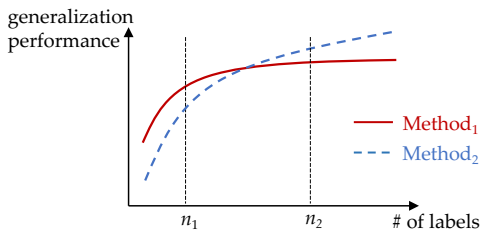
Fig. 1: Learning curves can inform NLP method selection given labeling budget. If the labeling budget is n_1, then Method$_1$ is preferred. If the labeling budget increases to n_2, then Method$_2$ is preferred.

ous works all use very large data sets for fine-tuning, which are often on the order of hundreds of thousands and even millions of examples [13], [14]. In general, however, it is impractical to collect training data at such a large scale in the health domain, for reasons discussed above.

Given the high potential of BERT and the often small data in health NLP, it is natural to ask the following question: *can we fine-tune BERT on small health NLP data and still achieve superior performance?* On the one hand, BERT may hold the promise as it has been shown to perform well in many NLP tasks thanks to the unsupervised pretraining. On the other hand, BERT is itself a large complex model with a massive number of parameters, so to achieve high performance it may need a good amount of labeled data for fine-tuning.

In this paper, we answer the above question by conducting learning curve analyses of BERT and other models on a disease diagnosis task. As conceptually shown in Figure 1, a learning curve can be viewed as a "return-on-investment" curve, where the "investment" is labeled data, and the "return" is a model's generalization performance on test data. Learning curves allow us to compare the performance of different models given different labeling budgets. They can also show which model will improve faster if we invest more labels. Such a comparison is especially relevant when the labeling cost is high, as in health NLP task scenarios.

The learning curve analysis reveals a series of interesting and informative findings, as summarized below:

- BERT is able to achieve superior performance even when fine-tuned on a handful of (but more than one) labeled documents per class.
- BERT's prior knowledge can effectively compensate for the lack of training data in most cases, but simple linear models are still worth considering when the amount of training data is extremely limited and not expected to increase any time soon. In the extreme case where each class has only one or two labeled documents, BERT could be outperformed by models using carefully engineered sparse bag-of-words features.
- When more labeled documents start to become available, BERT demonstrates fast rate of performance gain, which allows it to quickly outperform other models by a significant margin. It shows that BERT's prelearned repre-

sentation enables it to extract the most information from each training example. In other words, if we modestly increase the labeling budget, BERT will give the highest return.

II. RELATED WORK

BERT in health domain. Lee et al. obtained BioBERT by taking Google's pretrained BERT model and continuing the pretraining tasks (masked language modeling and next sentence prediction) on large-scale biomedical literature [15]. The use of domain-specific texts enabled BioBERT to out-perform BERT on certain biomedical NLP tasks. Alsentzer et al. [16] further added clinic texts to continue the pretraining on the basis of BioBERT to get Clinical BERT. A closely related line of work was conducted by Peng et al., where BERT is fine-tuned on biomedical and clinical texts, and then applied to ten benchmarking tasks, including sentence similarity measurement, named entity recognition, relation extraction, document classification, and logical inference [17]. All the above works demonstrate the value of domain-specific pretraining when applying BERT on health domain tasks.

This paper studies BERT from another significant perspective, *i.e.* its generalization performance when fine-tuned on small training data. To the best of our knowledge, there has been no previous work that studies the performance of BERT when the size of training data starts from very small. Instead, researchers often use learning curves to demonstrate the enormous learning capacity of deep learning models when training data size scales up exponentially [18].

Clinical text classification. Stanfill et al. conducted a systematic literature review of clinical coding and classification systems [19]. Recent works on clinical text classification studied various application scenarios, including smoking status identification [20], obesity prediction [21], [22], online patient forum moderation [23], cancer patient sentiment classification [24], vaccine adverse events detection [25], etc. These works above are all based on English texts. Zhong et al. [26] applied nearest neighbor classifier to identify the disease category based on patient disease description in Chinese. In this study, we predict the presence of a disease in clinical documents written in Chinese.

Incorporating existing knowledge. External knowledge has significant impact on machine learning performance. Besides pretraining model parameters using large unlabeled corpus, incorporating knowledge from ontologies (a.k.a. knowledge graphs or KGs) has also received attention. Garla et al. [27] utilized the relationship between clinical concepts in KG to improve feature selection. Yao et al. used UMLS entity embeddings in convolutional neural networks [22]. Li et al. used KG to derive additional knowledge features in rare disease classification [28]. Choi et al. [29] developed a graph-based attention model to represent words using node vectors learned from the ontology. Some studies [30], [31] suggest that incorporating KG into BERT also can bring some benefits.

III. METHODS

A. Data Description and Problem Formulation

We use two Chinese patient diagnosis corpora. The first corpus, HaoDaiFu, contains 51,374 patient records categorized into 805 diseases. Each document contains the symptom description submitted by a patient to Haodf.com, the largest Chinese online platform that connects patients to doctors. These patients have been previously diagnosed by a clinician, and now come to the platform for further consultation. The second corpus, ChinaRe, contains 86,663 patient records categorized into 44 disease categories. Each document contains the symptom description of a patient written by an insurance professional in ChinaRe, which is one of the largest reinsurance groups in China. The diagnoses were determined by a clinician and sent to the insurance company. Table I summarizes basic statistics of the two corpora. Jieba package was used for Chinese word segmentation [32].

TABLE I: Corpora statistics.

	HaoDaiFu	ChinaRe
# of documents	51,374	86,663
# of diseases	805	44
# of rare diseases	89	5
vocabulary size	59,879	41,087
average # of words/doc	27	30

Problem formulation. The task of patient diagnosis can be formulated as a text classification problem: to assign a disease label given the narrative description of a patient's symptoms. Accurate disease diagnosis is an important task towards computer-assisted patient triage and risk stratification. We aim to study the performance of different classification models (especially comparing BERT to other models) when provided an increasing amount of training data.

B. Compared Algorithms

In this section, we describe classification models we include in comparison. We broadly classify these models into two families according to their feature representation techniques.

1) Classifiers using sparse feature representations: We first consider the most common baseline of text classification – a linear classifier using bag-of-words features (and its variants). Although simple, such a model offers two advantages in handling small training data. First, a regularized, sparse linear classifier does not overfit as easily as complex models, therefore delivering stable performance. Second, the simple model allows relatively straightforward ways of incorporating prior knowledge into its feature representation.

BOW. This is a support vector machine classifier using TFIDF-weighted bag-of-words (BOW) features and linear kernel, trained with L_2 regularization.

BOW_EXP. This model enhances the feature representation of BOW with feature selection and synonym **exp**ansion techniques. The basic idea is to emphasize class-indicative features in a document if that document contains such a feature or its synonyms. It takes the following steps:

- A feature selection algorithm is used to rank the relevance of each unigram feature in the classification task.
- Each unigram feature w is associated with a class c if c has the largest $p(c|w)$ in training data. For each class, we select k highest ranking features according to the feature selection metric. The union of all selected features are denoted as F.
- For each word u in a document d, we compute its vector similarity to the vector of each $w \in F$ in a word embedding space. If cosine similarity $\cos(\vec{u}, \vec{w}) \geq t$, we increment the count of w in d by 1 before computing the TFIDF transformation. The step conceptually adds a new word w into d.

The above algorithm is a hybrid of feature selection and feature expansion [33]. Instead of discarding unselected features (which may still be useful), it increases the weights of selected features in each document. The method is inspired by the distributional prototype features proposed by [34] and later applied in clinical NLP [35].

BOW_EXP_KG. This model refines BOW_EXP by using knowledge graph (KG)-enhanced word vectors. A knowledge graph can be viewed as a semantic network, where entities (words and phrases) are nodes and relations between concepts are edges. We employ the LINE network embedding algorithm to learn low-dimensional word vectors that preserve knowledge in the semantic network [36].

2) Classifiers using dense feature representations: Recent text classification models represent words as dense feature vectors. One advantage of this approach is that pretrained word vectors can be easily used by such models, conveniently transferring semantic knowledge learned elsewhere to the current task. Here we consider two representative models using word vectors: the continuous bag-of-words model and long short-term memory networks.

CBOW. This is a linear-kernel support vector machine classifier that represents a document as the average of its words' vectors. It is also known as continuous bag-of-words (CBOW) [37], as conventional bag-of-words representation can be viewed as an average of one-hot word vectors. The word vectors are the same as in BOW_EXP and fixed in the training process.

CBOW_KG: This model refines CBOW by using KG-enhanced word vectors as used in BOW_EXP_KG.

LSTM: This classifier uses long short-term memory networks (LSTM) to process the document as a word sequence. The model's word embedding layer is initialized with the same word vector as in BOW_EXP and fine-tuned in the training process.

LSTM_KG: This model refines LSTM by initializing the word embedding layer with KG-enhanced word vectors as used in BOW_EXP_KG and CBOW_KG. The word embedding layer is fine-tuned in the training process.

3) Bidirectional Encoder Representations from Transformers (BERT): The crucial difference between BERT and recurrent neural network models (such as LSTM) is its extensive use of multi-head attention mechanism [38]. BERT represents each

word by "paying attention to" all other words in the sentence or short document. Instead of processing tokens sequentially as in LSTM, BERT's multi-head attention can process all tokens in parallel. This mitigates the gradient vanishing problem when capturing long-range dependencies between words. As a result, BERT can efficiently recognize the dependencies between labels and words as well as among words themselves.

BERT: We configure the Chinese BERT-base model released by Google[2] to perform multiclass classification tasks. We have not seen Chinese equivalent of BioBERT or ClinicalBERT. Nonetheless, the primary goal of our study is to compare BERT with other non-BERT classification models on small training data. It suffices to use a BERT-model pretrained on general domain texts.

C. Implementation Details

The support vector machine classifier was implemented using Python scikit-learn package. To determine the best regularization strength C for SVM models, we performed grid search over $\{10^{-3}, 10^{-2}, 10^{-1}, 10^0, 10^1, 10^2\}$ on a development set. We set $C = 1$ as it consistently delivered the best result (performance metric discussed below).

We explored various feature selection algorithms used in BOW_EXP and BOW_EXP_KG. These include chi-square (χ^2), information gain, and bi-normal separation [39] in our pilot study. We selected the χ^2 method as it delivers the best performance on development set. We select $k = 20$ features for each class.

In BOW_EXP and BOW_EXP_KG, the threshold of cosine similarity was set to $t = 0.9$ after searching over $\{0.7, 0.8, 0.9\}$ on development set.

We used 256-dimensional word vectors pretrained on a large-scale Chinese text corpus [40] in BOW_EXP_KG and CBOW_KG.

To learn KG-enhanced word vectors, we derive a semantic network from a general Chinese knowledge graph, CN-DBpedia [41]. It contains 16.8 million entities and 223 million relations and is publicly available.[3]

We used the LINE network embedding algorithm to fine-tune word vectors using the massive semantic network above. It was configured to learn from secondary-order proximity. We performed grid search for LINE's hyperparameters on a development set. These include (the best setting is underlined): negative edge sampling rate $\{5, \underline{10}, 50, 100\}$, batch size $\{128, 256, 512, \underline{1024}, 2048\}$, and number of batches $\{50K, 100K, 150K, 200K, \underline{250K}, 300K\}$.

We used tensorflow/keras to implement deep sequence learning models, including LSTM, LSTM_KG, and BERT. For LSTM models, we used the recommended Adam optimizer and default learning rate (10^{-3}). We set the number of training epochs such that the loss on validation set stops decreasing. For BERT, we also used the recommended Adam optimizer and default learning rate decaying schedule. The number of

[2]https://github.com/google-research/bert
[3]http://kw.fudan.edu.cn/cndbpedia/download/

training epochs was set to 40 using the same procedure as LSTM models.

D. Evaluation Methodology

Performance metric. Viewing the classification of each individual disease (class) as a binary classification problem, results can be divided into True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). *Recall* measures the percentage of TPs among all documents that truly mention that disease; *precision* measures the percentage of TPs among all documents predicted to mention that disease. F_1 score is the harmonic mean of precision and recall, a metric that balances the two [42]. To measure the classification performance of a set of diseases, we use macro-averaged F_1. Formally, the metrics are calculated as follow:

$$F_1 = \frac{2 \times recall \times precision}{recall + precision} = \frac{2 \times TP}{2 \times TP + FP + FN} \ , \quad (1)$$

$$\text{macro-averaged } F_1 = \frac{1}{|C|} \sum_{i=1}^{|C|} F_{1,i} \ , \quad (2)$$

where C is the set of diseases (classes), and $F_{1,i}$ is the F_1 score of the i-th disease.

Train-test split. To reduce the variance of results due to a random train-test split, we average the results of 10 runs. In each run, we randomly split the corpus into 80% for training and 20% for test. To avoid the case where some classes do not appear in training or test set, the random split is applied on a per-class basis.

Learning curve. The results of evaluation metrics we mentioned above are displayed in plots of learning curves. Learning curves represent the generalization performance of the models produced by a learning algorithm, as a function of the size of the training set. In a plot of learning curve, x-axis represents the size of training set, y-axis represents the performance of model under an evaluation metric. In our study, we sample training sets from total training examples in fixed proportions: [10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90%, 100%].

We use Area Under Learning Curve (ALC) to summarize the learning progress of each model. The ALC metric is useful in comparing different learning algorithms especially when labeling budget is limited, as in the active learning setting [43]. A higher ALC means an overall higher performance across different training data sizes.

IV. Experimental Results

The learning curves of different algorithms on HaoDaiFu and ChinaRe corpora are in Figure 2, with their corresponding ALC metrics reported in Table II. On both corpora, BOW_EXP, BOW_EXP_KG, and BERT significantly outperformed the BOW baseline, and BERT significantly outperformed the BOW_EXP_KG method.

To further study the behavior of different algorithms when training data are extremely few, we plot the learning curves on statistically rare diseases that account for no more than

0.02% (2 in 10,000) of records in each corpus. There are 89 such diseases in HaoDaiFu and 5 in ChinaRe. In both cases, these extremely rare diseases have on average about 10 training documents. This translates to *one training document per disease* at 10% training data rate, representing the cases of extreme data scarcity. The corresponding learning curves are in Figure 3, ALC metrics reported in Table III. On Hao-DaiFu, BOW_EXP, BOW_EXP_KG, and BERT significantly outperformed the BOW baseline on all diseases, and BERT significantly outperformed the BOW_EXP_KG method on extremely rare diseases. Since the number of rare diseases in ChinaRe is too few, the above performance comparisons did not show significant differences.

Note that the classification performance on HaoDaiFu is overall lower than that on ChinaRe. In general, multi-class classification problem is difficult with a large number of classes. Here the HaoDaiFu corpus contains one order of magnitude more classes than ChinaRe (*c.f.* Table I), bringing substantial challenge to all methods.

TABLE II: Area under learning curve (ALC) for different methods aggregated over all diseases. Figure 2 plots the learning curves.

Method	HaoDaiFu (all 805 diseases)	ChinaRe (all 44 diseases)
BOW	0.4158	0.8534
BOW_EXP	0.4226*	0.8934*
BOW_EXP_KG	0.4254*	0.8940*
CBOW	0.2097	0.5817
CBOW_KG	0.2064	0.5714
LSTM	0.0213	0.6064
LSTM_KG	0.0377	0.6243
BERT	0.5020*†	0.9551*†

*: result significantly higher than BOW.
†: result significantly higher than BOW_EXP_KG.
(Fisher's randomization test, significance level $\alpha = 0.05$)

TABLE III: Area under learning curve (ALC) for different methods aggregated over extremely rare diseases (prevalence $\leq 0.02\%$). Figure 3 plots the learning curves.

Method	HaiDaiFu (89 rare diseases)	ChinaRe (5 rare diseases)
BOW	0.3044	0.8454
BOW_EXP	0.3056*	0.9058
BOW_EXP_KG	0.3115	0.9034
CBOW	0.1215	0.1945
CBOW_KG	0.1153	0.2136
LSTM	0	0
LSTM_KG	0	0
BERT	0.3795*†	0.9028

*: result significantly higher than BOW.
†: result significantly higher than BOW_EXP_KG.
(Fisher's randomization test, significance level $\alpha = 0.05$)

V. DISCUSSION

BERT delivers the highest learning curve when aggregated across all diseases. With a fraction of all training data (30%

on Haodaifu, and 40% on ChinaRe), BERT is able to outperform all other approaches trained on 100% training data. These results show that BERT not only can deliver the best performance but also requires less data for training compared to other methods. The outstanding performance partly comes from Transformer's multi-head attention mechanism, which allows BERT to learn long-distance dependency much more efficiently than previous deep sequence models. It is also partly due to the unique pretraining objective, which can incorporate the sequence information of text in two directions efficiently.

BOW gives a decent baseline performance. Its variants, BOW_EXP and BOW_EXP_KG, give consistent performance improvements. Supervised feature selection and synonym expansion effectively improve the feature representation of BOW baseline. BOW_EXP_KG only gives slightly higher performance than BOW_EXP. This indicates that semantic relation information in a knowledge graph is already largely captured by pretrained word vectors.

CBOW performs worse than BOW. Similar result was observed in [44]. Indeed, linear SVM aims to find hyperplanes in the feature space to separate classes. It is easier to achieve linear separation in the high dimensional sparse feature space (BOW) than in the low dimensional dense feature space (CBOW).

The performance of LSTM on Haodaifu is extremely low, but is not that bad on ChinaRe, and goes up sharply when training data increases from 10% to 40%. This huge difference reflects the model's requirement for a large quantity of training data. On average, there are 51 training documents per disease in Haodaifu, while 1,575 training documents per disease in ChinaRe. Because of the vanishing gradient problem, training LSTM models becomes extremely difficult when training data size is small and documents are relatively long. Adding prior knowledge through word embedding (LSTM_KG) has only limited benefit.

On the extremely rare diseases (when there is only 1 training document per disease), BERT is outperformed by BOW_EXP and BOW_EXP_KG. This happened on both Haodaifu and ChinaRe. The result shows that in situations where training data is extremely scarce, the traditional non-deep model with an appropriate feature construction strategy is able to compete with the current state-of-the-art deep models.

On rare diseases, the catastrophically low performance of LSTM models is not unexpected given its poor performance on all diseases. Again, the result suggests that a large amount of training data is needed to train LSTM models, even its word embedding layer has been pretrained.

Implication. Medical domain has accumulated a wealth of knowledge bases, in the form of standardized terminologies, research publications, clinical practice guidelines, and consumer-facing information portals. While these forms of knowledge can be easily used by humans, they cannot be directly used by machine learning models. This is because the internal representation of knowledge in these models is fundamentally different from that of human knowledge. The primary way of transferring knowledge into these models
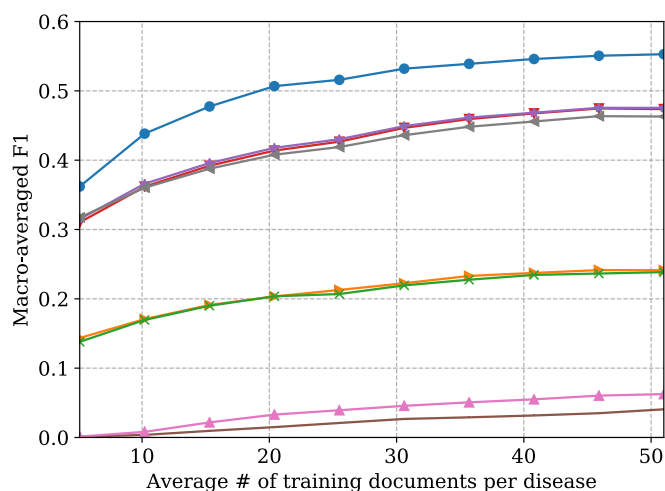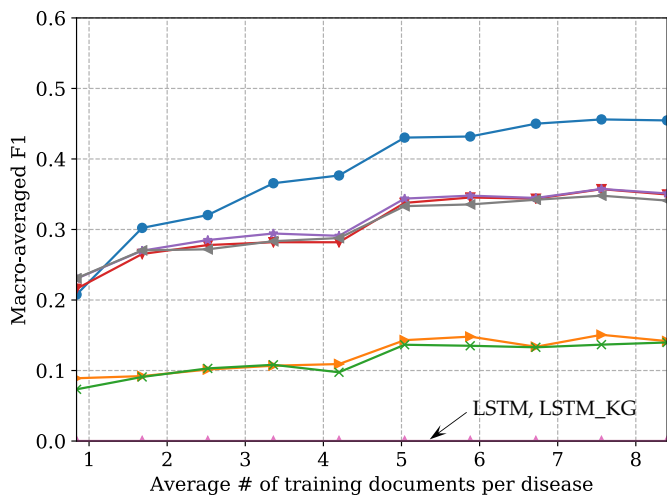
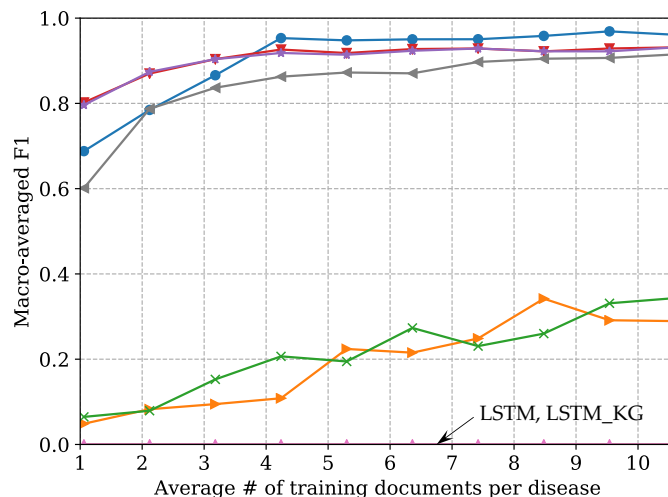(a) Learning curves for all 805 diseases in HaoDaiFu
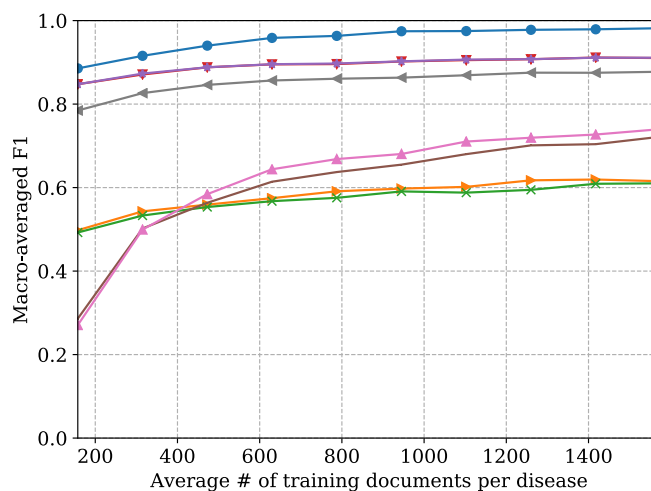
(b) Learning curves for all 44 diseases in ChinaRe

Fig. 2: Learning curves of compared algorithms averaged across all diseases in the two corpora.

(a) Learning curves for 89 rare diseases in HaoDaiFu

(b) Learning curves for 5 rare diseases in ChinaRe

Fig. 3: Learning curves of compared algorithms averaged across very rare diseases (prevalence $\leq 0.02\%$) in the two corpora.

is through well-formulated prediction tasks expressed in the form of labeled examples. However, labeling cost is high in the medical domain, necessitating machine learning models to leverage medical domain knowledge. Over the years, researchers have been proposing various methods for instilling external knowledge into machine learning models, including carefully designed features, model architectures, auxiliary learning objectives, and pretrained model parameters.

Our study here shows that pretrained BERT models (and the broader family of pretrained deep Transformers) may offer an effective way of leveraging external knowledge learned from large-scale unlabeled data towards specific NLP tasks. Even a BERT model pretrained on general corpus is able to effectively help NLP tasks in the health domain. On the one hand, this is good news to the health NLP research community, as it can potentially free researchers from feature engineering when the training data is small and the labeling cost is high. Instead, the model can be continuously improved by pretraining on domain-specific and task-specific corpora [15], [16], [45]. On the other hand, these black-box models are difficult to interpret, therefore more research is needed to understand their vulnerabilities especially within the medical context, such as potential biases in learned representations [46].

VI. Conclusion

In this paper, we study whether BERT is still effective when it is fine-tuned with small training data. To answer this question, we conducted a learning curve analysis of BERT and other baseline models in clinical text classification tasks. The analysis showed that BERT remains the highest performing model even when each class has only a handful of training documents, and its performance improves the fastest when given more training documents. Simple linear classifiers using specially engineered bag-of-words features delivers stable and competitive performance, and it outperformed BERT when training documents are extremely few (one or two per class). Overall, the study shows that even though BERT is a massively complex model, it only takes very small (but not extremely small) training data to fine-tune a pretrained BERT model to outperform baseline approaches using the same data.

References

[1] Amazon Web Services, "Amazon sagemaker ground truth pricing," Accessed July 2020. [Online]. Available: https://aws.amazon.com/sagemaker/groundtruth/pricing

[2] A. Stubbs and Ö. Uzuner, "Annotating longitudinal clinical narratives for de-identification: The 2014 i2b2/uthealth corpus," *Journal of biomedical informatics*, vol. 58, pp. S20–S29, 2015.

[3] United States Department of Health and Human Services, "National organization for rare disorders (nord)," Last Updated June 23, 2020. [Online]. Available: https://www.nidcd.nih.gov/directory/national-organization-rare-disorders-nord

[4] European Commission, "Rare diseases," Accessed July 16, 2020. [Online]. Available: https://ec.europa.eu/health/non_communicable_diseases/rare_diseases_en

[5] I. Spasic and G. Nenadic, "Clinical text data in machine learning: Systematic review," *JMIR Medical Informatics*, vol. 8, no. 3, p. e17984, 2020.

[6] A. B. Wilcox and G. Hripcsak, "The role of domain knowledge in automating medical text report classification," *Journal of the American Medical Informatics Association*, vol. 10, no. 4, pp. 330–338, 2003.

[7] D. Demner-Fushman, J. G. Mork, S. E. Shooshan, and A. R. Aronson, "Umls content views appropriate for nlp processing of the biomedical literature vs. clinical text," *Journal of biomedical informatics*, vol. 43, no. 4, pp. 587–594, 2010.

[8] P. I. Dissanayake, T. K. Colicchio, and J. J. Cimino, "Using clinical reasoning ontologies to make smarter clinical decision support systems: a systematic review and data synthesis," *Journal of the American Medical Informatics Association*, vol. 27, no. 1, pp. 159–174, 2020.

[9] E. Zhang, Q. Thurier, and L. Boyle, "Improving clinical named-entity recognition with transfer learning," *Stud Health Technol Inform*, vol. 252, pp. 182–187, 2018.

[10] G. Crichton, S. Pyysalo, B. Chiu, and A. Korhonen, "A neural network multi-task learning approach to biomedical named entity recognition," *BMC bioinformatics*, vol. 18, no. 1, p. 368, 2017.

[11] R. L. Figueroa, Q. Zeng-Treitler, L. H. Ngo, S. Goryachev, and E. P. Wiechmann, "Active learning for clinical text classification: is it better than random sampling?" *Journal of the American Medical Informatics Association*, vol. 19, no. 5, pp. 809–816, 2012.

[12] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[13] A. Adhikari, A. Ram, R. Tang, and J. Lin, "Docbert: Bert for document classification," *arXiv preprint arXiv:1904.08398*, 2019.

[14] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman, "Glue: A multi-task benchmark and analysis platform for natural language understanding," *arXiv preprint arXiv:1804.07461*, 2018.

[15] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "Biobert: a pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 2020.

[16] E. Alsentzer, J. R. Murphy, W. Boag, W.-H. Weng, D. Jin, T. Naumann, and M. McDermott, "Publicly available clinical bert embeddings," *arXiv preprint arXiv:1904.03323*, 2019.

[17] Y. Peng, S. Yan, and Z. Lu, "Transfer learning in biomedical natural language processing: An evaluation of bert and elmo on ten benchmarking datasets," *arXiv preprint arXiv:1906.05474*, 2019.

[18] J. Hestness, S. Narang, N. Ardalani, G. Diamos, H. Jun, H. Kianinejad, M. Patwary, M. Ali, Y. Yang, and Y. Zhou, "Deep learning scaling is predictable, empirically," *arXiv preprint arXiv:1712.00409*, 2017.

[19] M. H. Stanfill, M. Williams, S. H. Fenton, R. A. Jenders, and W. R. Hersh, "A systematic literature review of automated clinical coding and classification systems," *Journal of the American Medical Informatics Association*, vol. 17, no. 6, pp. 646–651, 2010.

[20] Ö. Uzuner, I. Goldstein, Y. Luo, and I. Kohane, "Identifying patient smoking status from medical discharge records," *Journal of the American Medical Informatics Association*, vol. 15, no. 1, pp. 14–24, 2008.

[21] Ö. Uzuner, "Recognizing obesity and comorbidities in sparse data," *Journal of the American Medical Informatics Association*, vol. 16, no. 4, pp. 561–570, 2009.

[22] L. Yao, C. Mao, and Y. Luo, "Clinical text classification with rule-based features and knowledge-guided convolutional neural networks," *BMC medical informatics and decision making*, vol. 19, no. 3, p. 71, 2019.

[23] J. Huh, M. Yetisgen-Yildiz, and W. Pratt, "Text classification for assisting moderators in online health communities," *Journal of biomedical informatics*, vol. 46, no. 6, pp. 998–1005, 2013.

[24] D. C. Edara, L. P. Vanukuri, V. Sistla, and V. K. K. Kolli, "Sentiment analysis and text categorization of cancer medical records with lstm," *Journal of Ambient Intelligence and Humanized Computing*, pp. 1–17, 2019.

[25] T. Botsis, M. D. Nguyen, E. J. Woo, M. Markatou, and R. Ball, "Text mining for the vaccine adverse event reporting system: medical text classification using informative feature selection," *Journal of the American Medical Informatics Association*, vol. 18, no. 5, pp. 631–638, 2011.

[26] J. Zhong, X. Yi, D. Xuan, and Y. Xie, "Categorization of patient diseases for chinese electronic health record analysis: A case study," in *Industrial Conference on Data Mining*. Springer, 2018, pp. 162–172.

[27] V. N. Garla and C. Brandt, "Ontology-guided feature engineering for clinical text classification," *Journal of biomedical informatics*, vol. 45, no. 5, pp. 992–998, 2012.

[28] X. Li, Y. Wang, D. Wang, W. Yuan, D. Peng, and Q. Mei, "Improving rare disease classification using imperfect knowledge graph," *BMC Medical Informatics and Decision Making*, vol. 19, no. 5, p. 238, 2019.

[29] E. Choi, M. T. Bahadori, L. Song, W. F. Stewart, and J. Sun, "Gram: graph-based attention model for healthcare representation learning," in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2017, pp. 787–795.

[30] Z. Zhang, X. Han, Z. Liu, X. Jiang, M. Sun, and Q. Liu, "Ernie: Enhanced language representation with informative entities," *arXiv preprint arXiv:1905.07129*, 2019.

[31] W. Liu, P. Zhou, Z. Zhao, Z. Wang, Q. Ju, H. Deng, and P. Wang, "K-bert: Enabling language representation with knowledge graph," *arXiv preprint arXiv:1909.07606*, 2019.

[32] "Jieba chinese text segmentation," Accessed March 26, 2019). [Online]. Available: https://github.com/fxsjy/jieba

[33] E. Gabrilovich and S. Markovitch, "Feature generation for text categorization using world knowledge," in *IJCAI*, vol. 5, 2005, pp. 1048–1053.

[34] J. Guo, W. Che, H. Wang, and T. Liu, "Revisiting embedding features for simple semi-supervised learning," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 110–120.

[35] Y. Wu, J. Xu, M. Jiang, Y. Zhang, and H. Xu, "A study of neural word embeddings for named entity recognition in clinical text," in *AMIA Annual Symposium Proceedings*, vol. 2015. American Medical Informatics Association, 2015, p. 1326.

[36] J. Tang, M. Qu, M. Wang, M. Zhang, J. Yan, and Q. Mei, "Line: Large-scale information network embedding," in *Proceedings of the 24th international conference on world wide web*, 2015, pp. 1067–1077.

[37] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[38] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[39] G. Forman, "An extensive empirical study of feature selection metrics for text classification," *Journal of machine learning research*, vol. 3, no. Mar, pp. 1289–1305, 2003.

[40] J. Su, "Pretrained word2vector," Accessed Apr 03, 2017. [Online]. Available: https://kexue.fm/archives/4304

[41] B. Xu, Y. Xu, J. Liang, C. Xie, B. Liang, W. Cui, and Y. Xiao, "Cn-dbpedia: a never-ending chinese knowledge extraction system," in *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*. Springer, 2017, pp. 428–438.

[42] Wikipedia, "F1 score," Accessed March 26, 2019). [Online]. Available: https://en.wikipedia.org/wiki/F1_score

[43] I. Guyon, G. C. Cawley, G. Dror, and V. Lemaire, "Results of the active learning challenge," in *Active Learning and Experimental Design workshop In conjunction with AISTATS 2010*, 2011, pp. 19–45.

[44] J. Lilleberg, Y. Zhu, and Y. Zhang, "Support vector machines and word2vec for text classification with semantic features," in *2015 IEEE 14th International Conference on Cognitive Informatics & Cognitive Computing (ICCI* CC)*. IEEE, 2015, pp. 136–140.

[45] S. Gururangan, A. Marasović, S. Swayamdipta, K. Lo, I. Beltagy, D. Downey, and N. A. Smith, "Don't stop pretraining: Adapt language models to domains and tasks," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, Jul. 2020, pp. 8342–8360. [Online]. Available: https://www.aclweb.org/anthology/2020.acl-main.740

[46] H. Zhang, A. X. Lu, M. Abdalla, M. McDermott, and M. Ghassemi, "Hurtful words: quantifying biases in clinical contextual word embeddings," in *Proceedings of the ACM Conference on Health, Inference, and Learning*, 2020, pp. 110–120.