Assessment of DistilBERT performance on Named Entity Recognition task for the detection of Protected Health Information and medical concepts

Macarious Abadeer
School of Computer Science
Carleton University
Ottawa, Canada
macarious.abadeer@carleton.ca

Abstract

Bidirectional Encoder Representations from Transformers (BERT) models achieve state-of-the-art performance on a number of Natural Language Processing tasks. However, their model size on disk often exceeds 1 GB and the process of fine-tuning them and using them to run inference consumes significant hardware resources and runtime. This makes them hard to deploy to production environments. This paper fine-tunes DistilBERT, a lightweight deep learning model, on medical text for the named entity recognition task of Protected Health Information (PHI) and medical concepts. This work provides a full assessment of the performance of DistilBERT in comparison with BERT models that were pre-trained on medical text. For Named Entity Recognition task of PHI, DistilBERT achieved almost the same results as medical versions of BERT in terms of $F1$ score at almost half the runtime and consuming approximately half the disk space. On the other hand, for the detection of medical concepts, DistilBERT's $F1$ score was lower by 4 points on average than medical BERT variants.

1 Introduction

Clinical records play an important role in the discovery of disease treatment and the advancement of medical research (Jagannatha and Yu, 2016). The clinical text corpora used for research includes doctor's notes, clinical study reports and medical articles. There are several regulations that control the use and transfer of personal information such as General Data Protection Regulation (GDPR) in Europe, Personal Information Protection and Electronic Documents Act (PIPEDA) in Canada and Health Insurance Portability and Accountability Act (HIPAA) in the US. HIPAA Safe Harbor for example lists 18 attributes that can potentially identify an individual and dictates that all of them need to be de-identified before a dataset can be shared for secondary use such as research (HIPAA, 2015). One possible approach is the manual annotation and de-identification of clinical text. This approach is simply not feasible due to the high cost of experts manually annotating clinical documents (Friedrich et al., 2019). Due to the advancement of Natural Language Processing research, the de-identification of PHI was framed as a Named Entity Recognition (NER) problem that can be solved by deep learning techniques. This work fine-tuned a deep learning model on a medical corpus and assessed its quality in detecting PHI and medical concepts in comparison with models whose embeddings were generated from a medical corpus.

The paper is organized as follows: in the next section we review the state-of-the-art for solving NER tasks used in the detection of PHI. In Section 3 and 4 we define the problem and detail our methodology. In Section 5 we present our results and we finally conclude in Section 6.

2 Related Work

In a major breakthrough in NLP research, a simpler neural network architecture was introduced by Vaswani et al. (2017) called Transformers which is an attention-based mechanism. Its main premise was to do away with recurrence and convolution in neural networks. Self attention generates a representation by connecting different positions of a given sequence. Self attention is easier to parallelize and enables better understanding of long-range dependencies. Transformers enabled the introduction of Bidirectional Encoder Representations from Transformers (BERT) by Devlin et al. (2019). BERT allows the generation of representations utilizing context from both directions of a sequence. It consists of two steps: pre-training and fine-tuning. Pre-training is the unsupervised learn-

Proceedings of the 3rd Clinical Natural Language Processing Workshop, pages 158–167
November 19, 2020. ©2020 Association for Computational Linguistics

ing step to generate the representations. BERT was pre-trained on Wikipedia and BookCorpus. The pre-training step consists of two tasks: Masked Language Model (MLM) which masks a certain percentage of the input sequence and attempts to predict those missing tokens. The second pre-training task is Next Sentence Prediction (NSP) which was specifically added to help with tasks involving relationship between a pair of sentences such as Question Answering. The fine-tuning is the supervised learning portion where BERT is trained on custom datasets by the user for their respective tasks with little to no feature engineering required for a specific NLP downstream task. Further attempts have been made to improve on the original BERT such as RoBERTa introduced by Liu et al. (2019) which assessed the impact of different hyperparameters and concluded that training over longer sequences and removing NSP achieves better results. Other BERT variations were also pre-trained on medical domain corpora such as BioBERT (Lee et al., 2019), BlueBERT (Peng et al., 2019) and ClinicalBERT (Alsentzer et al., 2019) that were pre-trained on PubMed which contains biomedical research articles and MIMIC-III which contains doctors' notes from the intensive care unit admissions. The medical versions of BERT proposed higher $F1$ score performance when evaluated on biomedical tasks including NER.

BERT and its variations, however, require extensive computational resources to deploy in production environments. To address these limitations, DistilBERT was introduced by Sanh et al. (2019). The authors applied the concept of knowledge distillation to produce a lighter version of BERT that is 40% smaller, 60% faster and achieves 97% of the original BERT $F1$ score when measured on Question Answering task. It can also be deployed on lower power computing chips such as mobile devices to run predictions. Further studies were published to assess the performance of DistilBERT compared to other state-of-the-art models. In a study by Büyüköz et al. (2020), it compared DistilBERT's performance against ELMo on two text classification tasks. The first was a binary classification task of protest and non-protest news from English articles from local newspapers in India and China. The second task was a sentence classification task of movie reviews on Rotten Tomatoes. The authors concluded that DistilBERT generalizes better than ELMo while having similar $F1$ score. Wang et al. (2020) also used DistilBERT for a machine translation task to generate synthetic data to diagnose language impairment in children. DistilBERT achieved 5% and 15% higher $F1$ scores when compared with ELMo and Word2Vec respectively.

3 Problem Statement

Although BERT achieved state-of-the-art for a wide variety of NLP tasks, they are hard to train and deploy in a production environment as they require excessive computational power. For example, the original BERT took 4 days to pre-train on 4 TPUs. Furthermore, there are few limitations of using a non-medical corpus to train a model for medical tasks (Patel et al., 2017). There are medical-specific terms that do not usually exist in general corpora such as news or Wikipedia. There are other terms that mean something else in a medical context. The idea of training on domain-specific corpora was explored by Cengiz et al. (2019) where the authors pre-trained BERT on specific domains such as telephone conversations, travel guides, government records and fiction novels. This achieved higher performance for related tasks than the generic version of BERT.

While versions of BERT pre-trained on medical text are publicly available as pointed out in Section 2, these models share the same computational power limitations of the original BERT. For example, the pre-training of ClinicalBERT took 18 days on a single GPU.

There are no studies we could find as of date that fine-tuned and assessed the performance of DistilBERT on medical tasks such as NER of PHI in medical records. Although in the context of de-identification predictions performance is more critical than runtime, the resource limitations may pose a challenge for healthcare organizations to comply with privacy regulations. Especially if they need to generate pre-trained embeddings or incrementally fine-tune their models on new data frequently.

The question we attempt to answer through this paper is how DistilBERT performs when fine-tuned on medical corpora compared to medical pre-trained versions of BERT. Is it possible to achieve a comparable result to medical pre-trained BERT variations such as ClinicalBERT with a much lighter version such as DistilBERT?

4 Method

DistilBERT is based on the concept of knowledge distillation introduced by Hinton et al. (2015). The main characteristic of a machine learning model evaluation is how it performs on unseen data. While high-confidence predictions are picked during inference, there are useful information in low-confidence predictions that can help explain how well a model can generalize. Knowledge distillation is a compression algorithm that involves the transfer of such information from the main model, called the teacher, to a smaller distilled version, called the student. Further details on knowledge distillation are in the paper by Hinton et al. (2015). DistilBERT consists of the same two steps as the original BERT: pre-training, which in this case creates the student model and fine-tuning which uses the pre-trained student model to train on a custom dataset for a specific task. DistilBERT was pre-trained on the same datasets as the original BERT: BookCorpus and Wikipedia. The assessment approach was to use the pre-trained DistilBERT and fine-tune it on i2b2 2010 and i2b2 2014 datasets for NER and compare the results with ClinicalBERT (Alsentzer et al., 2019) and BlueBERT (Peng et al., 2019) that were both pre-trained on medical text. The comparison was done in terms of runtime and $F1$ score.

The `transformers` package developed by Hugging Face Co[1] was used for all the experiments in this work. Its developers are also the creators of DistilBERT and it hosts a wide variety of pre-trained BERT models including the ones mentioned in Section 2. The package is implemented in `python` and this work was implemented in PyTorch.

Throughout this paper, by 'training' we are referring to the supervised learning step that BERT and its variants call 'fine-tuning' in order to avoid confusion with hyperparameter tuning. By 'pre-training' we are referring to the unsupervised step that generates the embeddings.

4.1 Datasets

i2b2 2014 - PHI: A dataset compiled by the National Center for Biomedical Computing (NCBC) also known as i2b2: Informatics for Integrating Biology and the Bedside. It contains doctors' notes provided by Partners HealthCare System in Boston. The 2014 version has an annotated text of PHI labels. The raw data is an XML file with positions of the PHI labels.

In total, there are 23 different labels with the top 3 accounting to 69% of all label instances (DATE, DOCTOR, HOSPITAL) and the bottom 7 having insignificant counts accounting to near-zero percentages. Since it was shown by Sokolova (2011) that using granular entities for PHI achieves better de-identification results than binary classification of whether an entity is a PHI, we chose to use all the labels for NER classification instead of binary PHI/non-PHI classification.

i2b2 2010 - Concepts: This dataset is also compiled by NCBC. It is another NER task that is focused on the extraction of medical concepts from patient reports. Specifically, it extracts medical problems, treatments, and tests. This dataset was included to validate whether models pre-trained on general domain corpora perform poorly on detecting medical terms and if yes, how poorly. Furthermore, medical history contains rich information about patients that HIPAA (2015) advised can individually identify a person.

Access to both datasets was requested through the Department of Biomedical Informatics[2] at Harvard Medical School which is provided for free to researchers and students.

The BERT model and its variations including DistilBERT expect NER datasets to be in CONLL-2003 format introduced by Tjong Kim Sang and De Meulder (2003). It was designed for NER tasks. Every line contains the word, a space, and the label of the entity in BIO format: B indicates the beginning token of a label, I for inside a multi-token label, and O for a token outside the entities to predict. Sequences are separated by two empty lines.

In order to produce the training, development and testing datasets in CONLL format from the raw files, we used the same scripts[3] used by ClinicalBERT authors (Alsentzer et al., 2019).

BERT variants including DistilBERT have a hard limit on sequence length set to 512 tokens. Some sequences in the raw datasets exceeded that limit. Those longer sequences had to be further split to fit the different sequence length experiments. The script referred to in the `transformers` pack-

[1] `https://github.com/huggingface/transformers`

[2] `https://portal.dbmi.hms.harvard.edu`

[3] `https://github.com/EmilyAlsentzer/clinicalBERT`

age's documentation[4] was used for splitting longer sequences. Table 1 shows token and sequence count after pre-processing.

	i2b2 2010		i2b2 2014	
	tokens	seq.	tokens	seq.
train	126,111	14,511	425,566	45,641
dev	7,612	1,804	58,053	5,241
test	229,992	27,626	306,441	32,587

Table 1: Tokens and Sequence Count for i2b2 2010 and i2b2 2014 after pre-processing

The train/test split for both i2b2 2010 and 2014 was already done by i2b2. In order to compare with other papers that used the same datasets as baseline, the train/test split was not modified even though i2b2 2010's training dataset has fewer tokens than its testing dataset.

4.2 Training

The NER examples provided by the `transformers` package was used as a starting point for training and evaluation. The full list of parameters used is discussed in Section 4.4. Adam optimizer (Kingma and Ba, 2014), a replacement to the generic stochastic gradient descent, was used for computing the loss function. The optimizer is initialized with learning rate, weight decay as well as the Adam ε constant set to 10^{-8} to avoid division by zero in the Adam calculation when the gradient approaches zero. A learning schedule was setup to dynamically modify the learning rate during training. The learning rate linearly increases during a phase of "warmup" steps, then linearly decreases after the warmup period. This is done because early on during training the model is far from convergence therefore updating the weights does not need to happen frequently. For every epoch in training, the loss is calculated, optimizer and scheduler steps incremented, model evaluated on the development set, and checkpoint is saved to disk.

4.3 Evaluation

The evaluation uses `seqeval.metrics` package to calculate precision, recall and $F1$ score. A classification report was also produced to display the scores for every label as well as the micro and macro average across labels for all 3 metrics. The classification report calculates the individual label scores using instances of the labels. For example, it does not calculate the scores for `B-DATE` and `I-DATE` individually but for the whole `DATE` label.

We ranked the best run based on micro average $F1$ score followed by recall if there's a tie in $F1$. In the context of de-identification, high recall is more critical since incorrectly annotating a non-PHI as a PHI token is less damaging than the opposite; or "leaking" personal information.

4.4 Experiments

The experiments were run on a GeForce GTX 1080 Ti, with 6 virtual cores, 64 GB of memory, 126 GB in hard drive storage and running Ubuntu 18.04.

The following are the different models that were experimented with:

distilbert-base-cased: DistilBERT English language model distilled from the cased version of Toronto BookCorpus and English Wikipedia.

distilbert-base-uncased: DistilBERT English language model distilled from the lowercase corpus version of distilbert-base-cased.

For the comparison with BERT variants pre-trained on medical corpus we used the following models:

BlueBERT Formerly known as NCBI BERT. A pre-trained version of BERT on uncased PubMed abstracts and MIMIC-III notes (Peng et al., 2019).

BioClinicalBERT: Also known as Clinical-BERT (Alsentzer et al., 2019). Another implementation of PubMed+MIMIC-III BERT which also included a hospital discharge summary corpus but pre-trained on cased text.

Both the cased and uncased versions of DistilBERT models are listed since they produced significantly different results. This is also required for a direct comparison since ClinicalBERT used a cased corpus while BlueBERT used an uncased one. Therefore, when comparing with ClinicalBERT, the cased version of DistilBERT was used. While when comparing with BlueBERT, the uncased version was used.

In total, 40 experiments were run to choose best-performing training parameters based on the highest micro average $F1$. For maximum sequence lengths, experiments ranged from using 128 to maximum allowed of 512. In terms of batch sizes,

[4] `https://github.com/stefan-it/fine-tuned-berts-seq`

16 and 32 were experimented with. Using a high maximum sequence length with a high batch size, however, resulted in out of memory issues. Therefore, 32 batch size was only used with maximum sequence length up to 256 while a batch size of 16 was used for higher maximum sequence lengths. All the experiments ran for 3 training epochs except one experiment ran for 2 epochs on the i2b2 2014 dataset to match the parameters reported by Alsentzer et al. (2019) for ClinicalBERT. The full range of parameters used in the experiments are shown in Table 2. The rest are all the defaults built-in the `transformers` package.

Parameter	Values
max. seq. length	$\{128, 150, 256, 300, 512\}$
batch size	$\{16, 32\}$
learning rate	5×10^{-5}
training epochs	$\{2, 3\}$
lowercase corpus	$\{True, False\}$

Table 2: Parameters experimented with

5 Results

We can draw the following insights from the results presented in Table 3. In terms of micro average $F1$ score, the performance gap between DistilBERT and its medical variants were dataset-specific. For the detection of PHI using i2b2 2014, DistilBERT scored within 0.5% of its clinical variants. It scored 0.56% higher than BlueBERT but 0.45% lower than ClinicalBERT. While for the detection of medical terms using i2b2 2010, the medical variants of BERT achieved 5% higher $F1$ score on average than DistilBERT. These results show that for the context of de-identification, using DistilBERT does not suffer in performance. This can be attributed to the generic nature of PHI labels such as dates, names and addresses that exist in general-domain corpus such as Wikipedia. While in the context of detecting medical terms, using a compressed model such as DistilBERT can result in significantly lower score than medically pre-trained models. Overall, the cased version of the models achieved $F1$ score of 6.82% higher on average than the uncased versions regardless of the dataset. This performance gap can be attributed to the importance of case information to the NER task according to BERT documentation[5].

Another aspect of the results we were interested in was runtime. As mentioned in Section 3, the original BERT is heavy to use requiring significant computing resources. DistilBERT runtime was 43% faster on average than the medical variants of BERT. It also produced a model that was consistently 60% smaller in size than BlueBERT and ClinicalBERT.

The per-label performance for all models and both datasets is shown in Appendix A. We can draw the following insights from the per-label comparison. As shown by support numbers column, the testing dataset for i2b2 2014 did not have significant counts for entities such as FAX, DEVICE and EMAIL therefore producing unpredictable $F1$ scores. This subsequently drove the average $F1$ lower. However, of the top 4 frequent labels (DATE, DOCTOR, PATIENT and HOSPITAL), DATE had the highest $F1$ score. On the other hand, HOSPITAL had the lowest $F1$ with significant support number (877 instances) achieving 48 $F1$ score which contributed to a lower micro $F1$ average for DistilBERT Uncased. On the other hand, DistilBERT cased model performed significantly better for the HOSPITAL label achieving 88 $F1$ score. As discussed earlier, casing features are important in the context of NER tasks. For English nouns, casing is particularly important. For example, hospital names are written with a capital first letter.

For i2b2 2010, since labels are all medical concepts, DistilBERT had trouble recognizing all 3 entities of treatment, problem and test achieving an $F1$ ranging from 78 to 82. For comparison, BlueBERT achieved 84-85 for all 3 entities and ClinicalBERT achieved 87.

The parameters that yielded the best performance out of all 40 runs are shown in Table 4. The parameters were dataset-specific but the same across all models.

In terms of relative performance, DistilBERT model's $F1$ score was, on average, 95% that of medically pre-trained BERT score for i2b2 2010 containing medical terms but on par for i2b2 2014 dataset. This result is 2 points lower than reported by Sanh et al. (2019) for question answering task. The performance degradation of using a distilled model is therefore task- as well as data-specific.

6 Conclusion

In this work the main contribution was a full performance assessment of DistilBERT in terms of

[5] https://github.com/google-research/bert

	Cased				Uncased			
	DistilBERT		ClinicalBERT		DistilBERT		BlueBERT	
	2010	2014	2010	2014	2010	2014	2010	2014
F1	83.48	94.85	87.51	95.38	79.56	86.44	84.05	86.05
Min.	19	60	34	102	18	31	34	50

Table 3: DistilBERT vs BERT Variants Results on i2b2 2010 & 2014 in terms of micro average $F1$ and runtime

	i2b2 2014	i2b2 2010
seq length	150	300
batch	32	16
epochs	3	
learning rate	5×10^{-5}	

Table 4: Parameters for best performing runs on i2b2 2010 and i2b2 2014

runtime and $F1$ score for the detection of medical concepts and PHI labels in medical records. DistilBERT was trained on a medical corpus using i2b2 2014 and i2b2 2010 datasets and compared the results with ClinicalBERT and BlueBERT; both are BERT variants that were pre-trained on medical corpora. For NER task of detecting PHI labels in medical records, DistilBERT achieved comparable results with twice the speed at approximately half the runtime. Its uncased version also performed slightly better in terms of $F1$ than BlueBERT. However, for detecting medical concepts such as problems, treatments and tests, DistilBERT's $F1$ score was lower by 5% on average than models such as BlueBERT and ClinicalBERT whose embeddings were generated from pre-training on medical corpus. Therefore, in the context of de-identification, using a distilled version of BERT such as DistilBERT produces very similar performance results at approximately 43% of the runtime compared to medically-trained BERT versions even when PHI labels are extracted from medical documents. Results shown here can guide the decision of adopting DistilBERT at healthcare organizations that need to frequently fine-tune their models on new medical data and use it for the detection of PHI labels. The reduced model size can also simplify the deployment process without performance degradation.

7 Future Work

Since DistilBERT achieved the same performance as medically-trained versions of BERT when detecting PHI labels even in medical context but suffered performance degradation when detecting medical concepts, future research can investigate and assess how DistilBERT performs on medical concepts if the student model was generated from a medical pre-trained teacher such as BlueBERT or ClinicalBERT. This involves pre-training DistilBERT using the same corpora as ClinicalBERT or BlueBERT in an unsupervised fashion to generate the embeddings. These embeddings can then be used for the fine-tuning step.

References

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

Berfu Büyüköz, Ali Hürriyetoğlu, and Arzucan Özgür. 2020. Analyzing ELMo and DistilBERT on socio-political news classification. In *Proceedings of the Workshop on Automated Extraction of Socio-political Events from News 2020*, pages 9–18, Marseille, France. European Language Resources Association (ELRA).

Cemil Cengiz, Ulaş Sert, and Deniz Yuret. 2019. KU_ai at MEDIQA 2019: Domain-specific pre-training and transfer learning for medical NLI. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 427–436, Florence, Italy. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Max Friedrich, Arne Köhn, Gregor Wiedemann, and Chris Biemann. 2019. Adversarial learning of privacy-preserving text representations for de-identification of medical records. In *Proceedings of* *the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5829–5839, Florence, Italy. Association for Computational Linguistics.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *stat*, 1050:9.

HIPAA. 2015. Guidance regarding methods for deidentification of protected health information in accordance with the health insurance portability and accountability act (hipaa) privacy rule. Accessed: April 11, 2020.

Abhyuday N Jagannatha and Hong Yu. 2016. Bidirectional RNN for medical event detection in electronic health records. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 473–482, San Diego, California. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Kevin Patel, Divya Patel, Mansi Golakiya, Pushpak Bhattacharyya, and Nilesh Birari. 2017. Adapting pre-trained word embeddings for use in medical coding. In *BioNLP 2017*, pages 302–306, Vancouver, Canada,. Association for Computational Linguistics.

Yifan Peng, Shankai Yan, and Zhiyong Lu. 2019. Transfer learning in biomedical natural language processing: An evaluation of bert and elmo on ten benchmarking datasets. In *Proceedings of the 2019 Workshop on Biomedical Natural Language Processing (BioNLP 2019)*, pages 58–65.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Marina Sokolova. 2011. Evaluation measures for detection of personal health information. In *Proceedings of the Second Workshop on Biomedical Natural Language Processing*, pages 19–26, Hissar, Bulgaria. Association for Computational Linguistics.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Yiyi Wang, Emily Prud'hommeaux, Meysam Asgari, and Jill Dolata. 2020. Automated scoring of clinical expressive language evaluation tasks. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 177–185, Seattle, WA, USA â†' Online. Association for Computational Linguistics.

A Per-label Performance

A.1 i2b2 2014

	precision	recall	f1	support
DATE	0.97	0.96	0.96	4988
DOCTOR	0.82	0.78	0.8	1915
PATIENT	0.79	0.75	0.77	881
HOSPITAL	0.56	0.41	0.48	877
AGE	0.98	0.98	0.98	764
MEDICALRECORD	0.96	0.96	0.96	422
CITY	0.76	0.74	0.75	260
PHONE	0.89	0.94	0.92	215
IDNUM	0.85	0.81	0.83	195
STATE	0.71	0.79	0.75	190
PROFESSION	0.75	0.7	0.72	179
ZIP	0.97	0.96	0.97	140
STREET	0.87	0.89	0.88	136
COUNTRY	0.6	0.24	0.34	117
USERNAME	0.99	0.96	0.97	92
ORGANIZATION	0.57	0.41	0.48	82
OTHER	0	0	0	13
DEVICE	0	0	0	8
FAX	0	0	0	2
EMAIL	0	0	0	1

Table 5: DistilBERT Uncased

	precision	recall	f1	support
DATE	0.99	0.99	0.99	4987
DOCTOR	0.95	0.95	0.95	1915
PATIENT	0.91	0.92	0.92	881
HOSPITAL	0.9	0.87	0.88	875
AGE	0.98	0.98	0.98	764
MEDICALRECORD	0.97	0.99	0.98	422
CITY	0.78	0.9	0.84	260
PHONE	0.93	0.97	0.95	215
IDNUM	0.8	0.88	0.84	195
STATE	0.88	0.8	0.84	190
PROFESSION	0.86	0.84	0.85	180
ZIP	1	0.96	0.98	140
STREET	0.95	0.97	0.96	136
COUNTRY	0.77	0.62	0.69	117
USERNAME	0.96	0.96	0.96	92
ORGANIZATION	0.7	0.55	0.62	82
OTHER	0	0	0	13
DEVICE	0	0	0	8
FAX	0	0	0	2
EMAIL	1	1	1	1

Table 6: DistilBERT Cased

	precision	recall	f1	support
DATE	0.97	0.96	0.97	4988
DOCTOR	0.82	0.75	0.79	1915
PATIENT	0.76	0.78	0.77	881
HOSPITAL	0.55	0.4	0.46	877
AGE	0.98	0.97	0.98	764
MEDICALRECORD	0.96	0.98	0.97	422
CITY	0.76	0.69	0.72	260
PHONE	0.9	0.95	0.93	215
IDNUM	0.87	0.77	0.82	195
STATE	0.67	0.77	0.72	190
PROFESSION	0.69	0.7	0.69	179
ZIP	0.96	0.96	0.96	140
STREET	0.88	0.9	0.89	136
COUNTRY	0.63	0.15	0.24	117
USERNAME	0.94	0.96	0.95	92
ORGANIZATION	0.53	0.44	0.48	82
OTHER	0	0	0	13
DEVICE	0	0	0	8
FAX	0	0	0	2
EMAIL	0	0	0	1

Table 7: BlueBERT

	precision	recall	f1	support
DATE	0.99	0.99	0.99	4987
DOCTOR	0.94	0.95	0.94	1915
PATIENT	0.93	0.93	0.93	881
HOSPITAL	0.88	0.86	0.87	875
AGE	0.98	0.98	0.98	764
MEDICALRECORD	0.97	0.99	0.98	422
CITY	0.76	0.85	0.8	260
PHONE	0.94	0.98	0.96	215
IDNUM	0.82	0.86	0.84	195
STATE	0.86	0.78	0.82	190
PROFESSION	0.8	0.87	0.83	180
ZIP	0.99	0.97	0.98	140
STREET	0.98	0.98	0.98	136
COUNTRY	0.68	0.48	0.56	117
USERNAME	0.94	0.96	0.95	92
ORGANIZATION	0.42	0.41	0.42	82
OTHER	0	0	0	13
DEVICE	0	0	0	8
FAX	0	0	0	2
EMAIL	0	0	0	1

Table 8: ClinicalBERT

A.2 i2b2 2010

	precision	recall	f1	support
problem	0.77	0.79	0.78	12592
treatment	0.79	0.79	0.79	9346
test	0.81	0.82	0.82	9226

Table 9: DistilBERT Uncased

	precision	recall	f1	support
problem	0.82	0.85	0.83	12593
treatment	0.82	0.84	0.83	9345
test	0.84	0.84	0.84	9226

Table 10: DistilBERT Cased

	precision	recall	f1	support
problem	0.83	0.84	0.84	12592
treatment	0.84	0.83	0.84	9346
test	0.84	0.86	0.85	9226

Table 11: BlueBERT

	precision	recall	f1	support
problem	0.86	0.88	0.87	12593
treatment	0.86	0.88	0.87	9345
test	0.86	0.87	0.87	9226

Table 12: ClinicalBERT