Big Data for Health Reproducibility Challenge – Reproducing Med7: NLP model for electronic health records

Avijeet Ranawat, Sidartha Rakuram
aranawat3@gatech.edu, sidartha@gatech.edu
Team #7 Github, Presentation Video

Abstract—Electronic health records are now ubiquitous but a majority of patient data remains in free-text form. We attempt to reproduce the Med7 model, a named-entity recognition model that recognises seven categories: drug names, route, frequency, dosage, strength, form, and duration. [1]

1 INTRODUCTION

In recent years, we observe a significant increase in patients' medical records being collected electronically. This provides a valuable opportunity for medical researchers to identify patterns in these datasets as we attempt to deliver improved healthcare for all patents worldwide.

Since a majority of these patients' medical records remain in free-text form, we aim to reproduce a named-entity recognition model that is able to parse the free-text inputs and identify the seven key medical concepts that are transcribed in the notes:

1) Drug names: e.g. aspirin, morphine, solumedrol
2) Route: e.g. gtt, drip, PO
3) Frequency: e.g. daily, twice a day, after meals
4) Dosage: e.g. One (1)
5) Strength: e.g. 1000 mg, 250-500 mcg/dose
6) Form: e.g. capsule, tablet, Aerosol Inhaler
7) Duration: e.g. two weeks, a month, one year

Identifying information from free-text is a topic that is studied heavily under "information retrieval", the sub-task of identifying entities within free-text is known as Named Entity Recognition (NER) [2].

The primary objective of Named Entity Recognition is to categorize these free-text records for future research. Since “manual processing of all patients’ free-texts records severely limits the utilization of unstructured data,” [1] this NER task allows to significantly reduce the cost of analyzing the overall dataset.

To train the model, a subset of the MIMIC-III dataset was used (part of the N2C2 2018 track 2) [3] dataset. In the original paper, the authors also extended that model to analyze its performance on UK mental health Clinical Record Interactive Search system CRIS data, which required additional tuning since these models do not generalize well outside of the contexts in which they are trained.

2 SCOPE OF REPRODUCIBILITY

As part of this reproducibility challenge, we aim to solve the following that was addressed in the original paper:

1. Build a named-entity recognition (NER) model that is robust to identify the 7 medical concepts indicated above.

However, due to resource and data availability constraints, the following hypotheses that were addressed in the original paper will not be addressed in our reproduction:

1. Active learning with human-in-the-loop approach to maximize the NER model’s accuracy. This is because we lacked a “human-in-the-loop” with sufficient medical experience to annotate additional data.
2. Attempting to generalize the model from the MIMIC EHR source domain to the UK mental health CRIS data. This is because access to UK-CRIS data could not be obtained since such data is restricted to NHS ‘approved researchers’ (including NHS staff and individuals from UK academic institutions).

3 METHODOLOGY

3.1 Model Descriptions

Model architecture:

We used the “en_core_web_trf” model in spaCy which uses a transformer-based architecture. In our initial runs we used the RoBERTa Base as the main architecture, later we replaced that with Bio_ClinicalBERT. These models are pre-trained on a large corpus of English text data using a masked language modeling. The Bio_ClinicalBERT is trained specifically on MIMIC III via next sentence prediction objectives. The en_core_web_trf model architecture consists of the following key components:

- Input Embedding Layer: This layer maps the input tokens to their corresponding embeddings, and this depends on the sequence size.
- Transformer Layers: Our model contains 12 transformer layers, each containing 12 attention heads. All these units are kept as default along with the hidden units and feed forward units.
- Output Layer: The output layer consisting of a linear layer maps the final hidden state to the 7 labels which we are detecting via NER.
- Activation Function : We are using the GELU (Gaussian Error Linear Units) activation function which performs well in transformer layers.

Training objectives:

Below are the different training objectives:

- Loss Function: Since our task was to find NER which contained multiple possible prediction labels we used cross-entropy loss.
- Optimizer: We used the Adam SGD optimizer since it adapts learning rate based on gradient.

Others:

We have eliminated the need to do pretraining on MIMIC III data. Since the original paper used a RoBerta base they had to pretrain on the entire MIMIC III to get the good results. We skipped this computation by using the pretrained Bio_ClinicalBERT which is already pre trained on the MIMIC III dataset.

3.2 Data Descriptions

We used the MIMIC III dataset [5], which contains the Electronic Health Records collected from patients admitted to the Beth Israel Deaconess Medical Center in Boston. MIMIC-III contains data from over 50,000 intensive care unit (ICU) admissions, including information on patient demographics, vital signs, laboratory test results, medications, procedures, diagnoses, and survival outcomes. The dataset also includes clinical notes such as physician progress notes, nursing notes, and discharge summaries.

We are using only a subset of the data including NOTEEVENTS.csv and PRESCRIPTIONS.csv for our analysis. An annotated version of the same is available in N2C2 (National NLP Clinical Challenges) dataset. Specifically the N2C2-2018 Task 2 contains annotated data of the clinical notes which contain the Drug names, Route, Frequency, Dosage, Strength, Form and Duration labels.

The n2c2 dataset that we used had a separate train and test dataset. Though validation data had to be created from training data by splitting it in the ratio of 80:20 such that 20% of the training data is used as validation set. In each of those dataset we had an equal number of .ann annotation files and .txt clinical notes files. Training dataset had 239, validation had 64 and testing had 202 .txt files in them along with similar named .ann files for them. Each line in the annotations file is tab separated with columns representing information such as the entity type, start and end offsets in the text data, along with any additional attributes or relationships. Labels in the dataset are imbalanced in the sense that Drug Names have around 6k annotations, while Form has only 600. All the other labels are around 1k to 1.5k each.

Dataset Links:

- MIMIC III : https://physionet.org/content/mimiciii-demo/1.4/
- N2C2 : https://portal.dbmi.hms.harvard.edu/projects/n2c2-nlp/

3.3 Computational Implementation and Requirements

In terms of softwares, the members with physical GPU access used VS Code, while others used colab notebooks with GPU runtime for development. We are using GitHub for Version Control. The RAM we used to reproduce the paper is 32GB along with a common GPU available on Mac and NVIDIA-SMI 525.85.12 GPU available in colab. With these configs, we were able to train and validate the model in approximately 2 hours.

3.4 Hyperparameters

Table below shows the various hyperparameters tweaked for the 3 models which we tried. Major effects were due to the learning rate and max epochs allowed parameter.

Table 1—Hyperparameters of the different models trained

Hyperparameters/ Models	initial_rate [LR]	dropout	max_steps	batch_size	max_epochs	hidden_width
A.	0.00005	0.1	20000	50	10	64
B.	0.0001	0.2	10000	25	15	128
C.	0.000001	0.3	50000	10	25	256

3.5 Code and Environment

We use Python as the primary programming language. The spaCy language model is used for the NLP related concepts like NER tagging and parsing. We utilized the pipeline creation feature of spaCy and included transformer, NER components to our model. Original paper provided only the resultant model via hugging face for further utilization. Codes to reproduce those models or other data processing were not given by the authors. Though we utilized other resources like these to know more about pipeline creation via spaCy [6].

GitHub Jupyter Notebook Link (access granted to our grader, Xiaocheng Chen): https://github.gatech.edu/srakuram3/med7-reproduction/blob/main/spacy_train.ipynb

4 RESULTS

The table below shows the precision, recall and F1 scores for the 3 different models we built to identify the various entities. We saw that due to the very small learning rate of 0.000001 model C could not learn much in the speculated max epochs of 25. So it showed a score of 0 in multiple labels. While model A and C were way more better right from the starting epochs. Model A also started overfitting after epoch 7. While the original paper also utilized human annotated gold dataset which we lacked due to unavailability of domain expertise, we were able to identify most of the labels and got 85+ F1 scores in most of them.

Table 2—The evaluation results of our models using the 202 annotated documents from N2C2

	Model A			**Model B**			**Model C**		
Entity	**Precision**	**Recall**	**F1 Score**	**Precision**	**Recall**	**F1 Score**	**Precision**	**Recall**	**F1 Score**
Drug	89.41	81.67	85.36	89.33	88.89	89.11	56.8	64.49	60.4
Route	91.37	80.66	85.68	92.07	90.07	91.06	91.45	11.89	21.05
Strength	81.88	80.68	81.28	88.55	93.69	91.05	46.96	47.65	47.31
Form	79.04	71.06	74.84	81.59	84.83	83.18	48.93	38.94	43.36
Dosage	91.24	64.51	75.58	86.09	86.06	86.07	0	0	0
Frequency	74.52	69.64	72	65	70.22	67.51	1.72	0.1	0.19
Duration	78.9	23.76	36.52	79.47	41.71	54.71	0	0	0

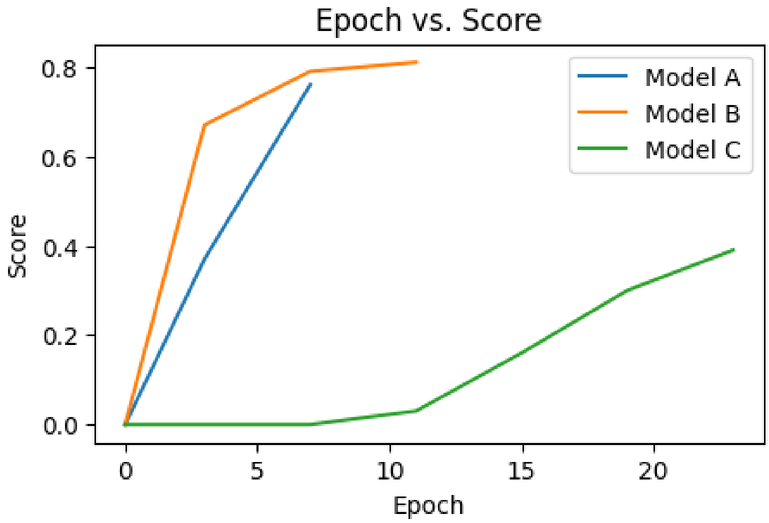


Figure 1—Results of the three different models

5 DISCUSSION

Based on our attempt above, we were able to reproduce the results obtained by Kormilitzin to detect the seven main categories of medical concepts from the 2018 n2c2 dataset. We achieved an overall score of F1=0.84 which was slightly lower than the paper's F1=0.893. As discussed in the Scope of Reproducibility section, we didn't explore evaluating the transferability of this model to the mental health records (CRIS) in the UK due to difficulty obtaining access to the UK CRIS dataset.

Overall, we can observe that our present model is getting an average F1 score of 84%. It is able to cross 90+ scores in tagging "Route" and "Strength", also we see 83+ scores in "Dosage", "Drug" and "Form".

Since the original authors did not share the reference code for how they implemented the models, we had to start from scratch in building the models based on the methodology highlighted in the section 3.2 of the original paper [1].

In the beginning, we faced a few challenges in identifying how RoBERTa fit into the overall model architecture, and we spent a significant amount of time trying to connect the two model architectures. When we determined that spaCy had a mechanism to set the transformer base architecture, it was a major breakthrough for our reproduction.

An aspect that helped our reproduction is the pre-compiled SpaCy model that the authors published on HuggingFace. While this didn't give us information on how to rebuild the model, we at least had a reference on how a "correct" model should operate and the performance of that model. This would allow us to compare how far we've come in building our model.

To allow easier reproducibility, we recommend that authors of papers provide a link to a GitHub repository with the model setup, training, and validation steps. We acknowledge that maintaining a GitHub repository does take additional effort to ensure compatibility with the latest versions of different software packages (like newer versions of Python, spaCy, TensorFlow, etc.).

6 REFERENCES

[1] A. Kormilitzin, N. Vaci, Q. Liu, and A. Nevado-Holgado, "Med7: A transferable clinical natural language processing model for electronic health records," Artificial Intelligence in Medicine, vol. 118, p. 102086, 2021, doi: 10.1016/j.artmed.2021.102086.

[2] C. D. Manning, P. Raghavan, and S. Hinrich, Introduction to Information Retrieval. Cambridge University Press, 2019.

[3] S. Henry, K. Buchan, M. Filannino, A. Stubbs, and O. Uzuner, "2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records," Journal of the American Medical Informatics Association, vol. 27, no. 1, pp. 3–12, Oct. 2019, doi: 10.1093/jamia/ocz166.

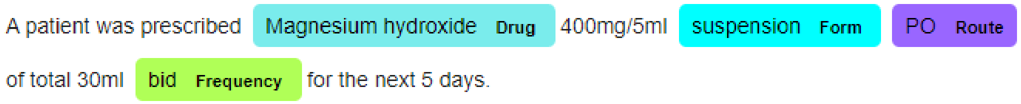
[4] E. Alsentzer et al., Publicly Available Clinical BERT Embeddings. 2019.

[5] A. Johnson, T. Pollard, & R. Mark (2019). "MIMIC-III Clinical Database Demo" (version 1.4). PhysioNet, doi: 10.13026/C2HM2Q.

[6] M. Honnibal, I. Montani, S. Van Landeghem, & A. Boyd, (2020). spaCy: Industrial-strength Natural Language Processing in Python. https://doi.org/10.5281/zenodo.1212303

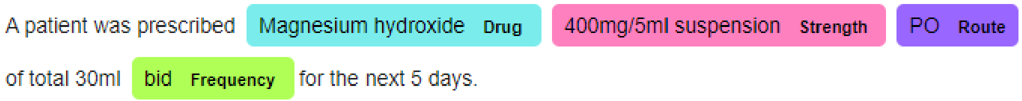
7 APPENDIX

7.1 Model A results



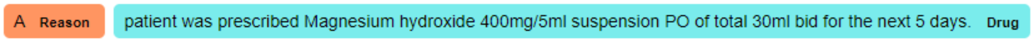
A patient was prescribed Magnesium hydroxide **Drug** 400mg/5ml suspension **Form** PO **Route** of total 30ml bid **Frequency** for the next 5 days.

7.2 Model B results



A patient was prescribed Magnesium hydroxide **Drug** 400mg/5ml suspension **Strength** PO **Route** of total 30ml bid **Frequency** for the next 5 days.

7.3 Model C results



A **Reason** patient was prescribed Magnesium hydroxide 400mg/5ml suspension PO of total 30ml bid for the next 5 days. **Drug**