Using Natural Language Processing to de-identify medical data

Abrar Majeedi1, Ashish Singh1

1University of Wisconsin-Madison, United States

1. Abstract

With the growth in data collection and the digitization of almost all the record keeping aspects of human life, tremendous developments have been made by researchers using that data. One field where there is no dearth of data is the field of medicine. Data ranging from test results to clinical notes to prescriptions, almost all of it is digitized and available. But the presence of patient identifiers like names, contact details and dates make it impossible to share the data unless very time consuming and intricate protocols of data sharing are followed to protect patient privacy. This has stifled the use of advanced data analysis and modern AI to solve some very important problems in the field of health care. One solution to this problem is to remove all the Personal Health Information (PHI) identifiers from data. The resulting data can then be shared freely without any concerns of patient privacy. In this project we develop and evaluate an approach to de-identify clinical text notes using current state-of-the-art technique from Natural Language Processing (NLP) called Bidirectional Encoder Representations from Transformers (BERT). We propose human interpretable evaluation measures and demonstrate state of the art performance against modern baseline models. All the code is open source and highly portable [1].

2. Introduction

In the current age of Artificial Intelligence, we have a lot of advanced data analysis and machine learning methods available that can solve problems considered impossible before. The only requirement is data. More that data better the models will perform. This is where the problems in domain of healthcare stand out. Data driven research in healthcare has a bottleneck in the data availability. A lot of data is collected but sharing it is very difficult due to the HIPAA restrictions. While availability of data would be great, but at the same time patient privacy is very important too. As a result, most of the NLP research is focused on the few freely datasets available. One of these datasets is MIMIC-III. Consequently, we have also used the same dataset to develop and test our model on same benchmarks as previous approaches.

In this study we develop and evaluate a BERT based text de-identification model which achieves state-of-the-art results for de-identifying clinical text. Since Johnson et. al. recently published their BERT based de-identification model [1], we try to replicate their results in this project. We are also making all the code available on github for reproducibility of results and portability[2].

3. Data

For this project we have used the MIMIC-III dataset [3] made available by Laboratory for Computational Physiology, MIT USA. This data set consists of a wealth of data related to 61,532 intensive care unit stays: 53,432 stays for adult patients and 8,100 for neonatal patients. We focused on the free text clinical notes made available in this dataset. The free text consists of about 3.8 billion characters. Due to computational limits, we selected a subset (500 million characters or 13.15%) of this dataset for our project. The data is thoroughly deidentified, but no annotations are available.

3.1 Train Test Split

The data is split into train and test sets with a ratio of 3:1, i.e. the model is trained on 75% of the data and the performance is measure on 25% of data, which was not seen by the model during training.

4. Preprocessing

This step was by far the most complex and time consuming in the entirety of this project. Since the MIMIC-III data set is de-identified, we were surprised to find that no annotation was made available. By annotation we mean, each word in the text should have been labelled as “Name” or “Age” or “Location” or “non-PHI” and so on. Ultimately, we decided to re-identify the data on our own. Some of the PHI like names and addresses have been replaced by placeholders of a uniform pattern which made it easier to identify which words are PHI. We replaced all the placeholders with realistic looking PHI. This results in a re-identified dataset, and at the same time we label all the words based on regex patterns and string matching.

**4.1 Re-identification**

We use regex and string matching to identify different PHI. After the PHI is identified, we replace the different place holders in the following manner:

1. First names: These are filled in by random sampling from a list consisting of 600 most common male and female names in the US.
2. Last names: These are filled in by random sampling from a list consisting of 600 most common last names in the US.
3. Hospitals: These are filled in by random sampling from a list consisting of 500 most popular hospital names in the US.
4. Contact numbers, Unique IDs, Dates and Age: These are filled using the python random module, with proper formatting (e.g. Phone numbers as xxx-xxx-xxxx and 1-2 digit ages, SSN as xxx-xx-xxxx and so on)
5. Addresses: These are filled in using fake but real looking addresses generated on [https://www.randomlists.com/random-addresses](https://www.randomlists.com/random-addresses?)

Once the data is re-identified and the corresponding labels for each word are recorded as well, it is finally ready to be used for the purpose of Named Entity Recognition, i.e. learning how to label each word in the text. In our case we limit the labels to:

Names, Contacts, Dates, Locations, Unique IDs (Insurance numbers, SSNs etc.) and finally non-PHI.

**4.2 Initial Tokenization**

We use the NLTK library in python to perform the initial tokenization, i.e. splitting the sentences into individual tokens (which at a higher level can be seen as splitting sentences into individual words). The corresponding labels are distributed over the resulting tokens if a word leads to more than one token.

5. Approach

5.1 BERT

A screenshot of a cell phone

Description automatically generatedDevlin et al. pre-train deep Bidirectional Transformers for language modeling [4] to create a model named BERT, which is capable of simultaneously contextualizing word embeddings using all flanking context. This model resulted in significant improvements on the state-of-the-art in NLP on several tasks.

5.2 Model

We used uncased SciBERT from the Allen-AI NLP repository for our project. It is a BERT model that has been pre-trained on scientific corpora. It is “uncased” because it does not take the case of letters into account. We did not use versions of BERT fine-tuned to clinical corpora, such as ClinicalBERT, because these are trained using deidentified text and are known to suffer weaker performance for the task of deidentification [3]. The tokens are further tokenized by piecewise tokenization, which is basically splitting words into sub words.

The model takes 8 inputs tokens at a time. These tokens are first converted to ids based on the vocabulary of the SciBERT model and then passed through the model. We feed the final hidden representation of BERT into a fully connected dense layer with one output for each *Figure 1*

entity type, including the “non-PHI” label. Figure1

shows the architecture of the model.

5.3 Training

The model is trained on a Google Cloud Instance with Tesla V4 GPU. We fine tune all the weights including those in the pre-trained SciBERT. We used Adam optimizer with a learning rate of 3 × , and epsilon of . The model is trained using CUDA 10.1 and PyTorch 1.3 with a batch size of 32 for 3 epochs. The training took 17 hours.

6. Results

Since we do not use the entire data set for our project, the results are not exactly comparable with previous methods. But we are sure that if we try our approach for the entire dataset, the results will remain close to what we have right now. Also, our results are all measured at token level. Table 1 summarizes the results. Also Figure 2 shows the ROC curves and Figure 3 shows the Precision-Recall curves for each class label.

Table 2 compares the results of our approach with the current State-of-the-art method and other approaches that have been tried for deidentification of clinical text notes.

|  |  |  |  |
| --- | --- | --- | --- |
| **Class** | **Positive Predictive Value (Precision)** | **Sensitivity (Recall)** | **F1-score** |
| *CONTACT* | 0.98336516 | 0.98241133 | 0.98288801 |
| *DATE* | 0.99244865 | 0.99275729 | 0.99260295 |
| *LOCATION* | 0.99569739 | 0.99507795 | 0.99538757 |
| *NAME* | 0.99179696 | 0.99013438 | 0.99096497 |
| *OTHER (non-PHI)* | 0.99934905 | 0.99934959 | 0.99934932 |
| *UNIQUE ID* | 0.96716848 | 0.97340426 | 0.97027635 |

Table 1. Summary of results

A screenshot of a cell phone

Description automatically generatedA screenshot of a social media post

Description automatically generatedA screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generated

A screenshot of a social media post

Description automatically generatedA screenshot of a cell phone

Description automatically generated

Figure 2 ROC Curves for Individual classes (one-vs-all)

A screenshot of a social media post

Description automatically generatedA screenshot of a social media post

Description automatically generated

A screenshot of a social media post

Description automatically generatedA screenshot of a social media post

Description automatically generated

A screenshot of a social media post

Description automatically generatedA screenshot of a cell phone

Description automatically generated

Figure 3. Precision-recall curves for Individual classes (one-vs-all)

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **PPV** | **Recall** | **F1** |
| Ours | 0.98830428 | 0.98885579 | 0.98857819 |
| Johnson et. al. | 0.9561 | 0.9561 | 0.9561 |

Table 2. Comparison of our results with the current state-of-the-art

7. Conclusion

We have demonstrated and replicated the results of Johnson et. al [1]. We believe that the difference in our results and their results is primarily because of using a smaller dataset and because we report the performance metrics at token level.

We can thus conclude that BERT based de-identification methods achieve great results and can be used in combination with other automated de-identification methods e.g. handcrafted regex to make a system that can reliably remove all PHI from a dataset, thus making sharing of clinical text data between researchers much easier. This will revolutionize data-driven research in the field of medicine.

8. References

1. Deidentification of free-text medical records using pre-trained bidirectional transformers. Johnson A., Bulgarelli L, Pollard T CHIL (2020)
2. https://github.com/abrarmajeedi/BERT-based-medical-text-de-identification.
3. MIMIC-III, a freely accessible critical care database. Johnson AEW, Pollard TJ, Shen L, Lehman L, Feng M, Ghassemi M, Moody B, Szolovits P, Celi LA, and Mark RG. Scientific Data (2016).
4. BERT: Pre-training of Deep Bidirectional Transformers for Language. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. (2018)