Electronic Health Records (EHR) Extraction Using Various NLP Models

Aman Mukhopadhyay

***Abstract:** Information recorded in electronic medical health records, clinical reports, and summaries has the possibility of revolutionizing health-related research and its corresponding industry. EMR data can be used for epidemiological studies, disease registries, data banks, drug safety surveillance, clinical trials, and healthcare audits. With the rapid adoption of electronic health records (EHRs), it is desirable to harvest information and knowledge from EHRs to support automated systems and to enable secondary use of EHRs for clinical and translational research; thereby increasing efficiency. One critical component which is predominantly used to facilitate the secondary use of EHR data is information extraction (IE) task, which automatically extracts and encodes clinical information from a given text. Now, a natural language processing model (NLP) focuses on "developing computational models for understanding the interaction between data science and language". In the clinical domain, researchers have often used NLP systems to identify clinical syndromes and common biomedical concepts from imaging data, radiology reports, discharge summaries, problem lists, nursing documentation, drug reviews, and medical education documents. These data can help doctors determine patients' health condition(s) including diagnostic information, procedures and tests performed, treatment results, drugs administered, and more.*

Therefore, we hope to gain some insights and develop strategies to improve the utilization of these NLP systems in the clinical domain. We hope to provide a vision for addressing the existing data challenge(s) in this domain. For this, we would look at the various models that have been used/published over the years and test them for their attributes including effectiveness, accuracy, precision, etc. We believe that adding a probabilistic graphical model framework for structured output prediction would further improve the performance of our system. This experiment remains our future work.

***Keywords:** Electronic Health Records, Machine Learning, Models, Natural Language Processing, Disease, Information Extraction, Processing Data.*

I. INTRODUCTION

EHRs are huge free-text data files or datasets that are documented by healthcare professionals, e.g.: clinical notes, imagery results, discharge summaries, or lab reports. Finding specific information from this data is obviously time-consuming since the data is unstructured and there may be multiple such records for a single patient.

So, NLP techniques can be used to make this data structured, and quickly find information whenever needed, thereby saving the time needed from otherwise completing these mundane tasks.

In this project, we aim to build a tool that would automatically structure this data into a format that would enable doctors and other healthcare professionals to quickly find the information that they need. Specifically, we aim to build a Named Entity Recognition (NER) model that would recognize entities for a case event such as drug strength, administered time, duration, frequency, adverse drug event (ADE), reason for taking the drug, route, and form. Further, the model can also recognize the relationship between drugs and every other named entity as well [1]. This would allow healthcare professionals to not only look at individual entities but also all the relationships between them (if present). Moreover, this would also allow them to easily find out the relationships between a drug and ADEs so that such drugs can be monitored carefully [4].

Subsequently, the final goal of this project would be to build an API where healthcare professionals could potentially send EHR data and the API would return character and/or data ranges for each annotation so they can be highlighted in the original data. In other words, it would return structured JSON-format data that includes separately labeled data for medication history and discharge medications. The highlighted annotations could be useful when a healthcare professional wants to see important information along with other details in the EHR. The structured information can then be used to store the data for quick reference in the future.

II. MATERIAL AND METHODS-OVERVIEW OF THE PROPOSED SYSTEM

A. Named Entity Recognition (NER)

To perform NER on a given dataset, three different models are built. A rule-based model is built as a baseline and two machine learning models are then subsequently built.

i. Rule-Based Model

To establish a unanimous baseline, a traditional dictionary, and regular expression-based NER model is used. A regular expression is written to find the dosage entity, which would find any number followed by "mg" or "mcg". For all other entities, the data is split into 80% train data and 20% test data. The training data is used to create a dictionary of each entity, so if the same entities appear in the test data, it would classify it as the corresponding entity.

ii. BiLSTM + CRF

Just a BiLSTM network is enough to classify each token into various entities along with its class (i.e. B: beginning or I: inside). Since the outputs of BiLSTM of each word are the label scores, we can select the label that has the

Manuscript received on 14 March 2024 | Revised Manuscript received on 25 October 2024 | Manuscript Accepted on 15 December 2024 | Manuscript published on 30 December 2024.

*****Correspondence Author(s)

Aman Mukhopadhyay*, Department of Computer Science and Engineering, VIT University, Vellore (Tamil Nadu), India. Email ID: amanmukherjee073@gmail.com, ORCID ID: 0000-0001-7494-0562

highest score for each word. By this scheme, we may end up with invalid outputs, for e.g.: I-Drug followed by I-ADE or B-Drug followed by I-ADE.

iii. BioBERT

The output of each token from the BERT model is passed through a fully connected neural network with a SoftMax layer at the end that classifies that token to an entity. The entities here would be in IOB format, for example, B-DRUG and I-DRUG would be treated as separate entities. This entire model is called BERT for token classification, and its architecture is available in Python's transformers library. BioBERT is a pre-trained BERT model, which is trained on a medical corpora of more than 18 billion words. Since it has a medical vocabulary and is trained in biomedical data, we chose this model for finetuning our dataset.

B. Data Preprocessing

Usually, the EMR databases are composed of a variety of heterogeneous data sources, and so the data retrieved from the EMR database is diverse, incomplete, and redundant, which in turn, will affect the final mining result [2] to a great extent. Therefore, the EMR data must be appropriately preprocessed to ensure that the data is accurate, relevant, complete, and consistent, and has protected privacy. The process of data preprocessing includes data cleansing, data integration, data transformation, data reduction, and privacy protection.

C. Data Cleaning

The EMR data, which is incomplete, noisy, and inconsistent, can be cleaned by filling in the defaults, smoothing noise, and correcting data inconsistencies. When gathering EMR data, some data attributes may be lost due to manual errors and system failure. For default data, there are several ways around this. We could ignore missing data, manually fill default values, use attribute averages, fill defaults with the most likely values, or retrieve other data sources. When the missing value has a small influence on the processing process, the missing data is usually ignored. Furthermore, when the missing data attribute exists in other data sources, the data source should be retrieved.

D. Data Integration

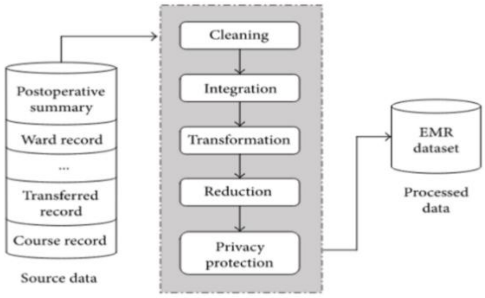
At the data integration stage, the data stored in different data sources needs to be consolidated and assimilated thoroughly. But here comes the challenge of dealing with heterogeneous data and its redundancy. Through data integration, the accuracy and speed of data mining can be improved. In a nutshell, if an attribute can be derived from other attributes, then the attribute is redundant and should be cleaned up. Redundancy is mainly reflected in the repeated records of data attributes or inconsistencies in the way of attribute expression.

E. Data Transformation

Data transformation refers to the conversion of a dataset into a unified form suitable for data mining. Suitable methods include smoothing noise, data aggregation, and data normalization. According to the direction and target of data mining, the data transformation method filters and summarizes EMR data. Data analysis can be more efficient by having directional, purposeful data aggregation. To avoid the dependency of the data attributes on the measurement units, data should be normalized to make the data fall into smaller common spaces, to make it more readable.

F. Privacy Protection

Compared with traditional paper medical records, the application of EMR has greatly promoted the development of medical care, but it has also brought a lot of privacy and security problems. EMRs contain sensitive information about the patients, and it can be very serious if gone into the wrong hands. So, to mitigate these issues, we can adopt different ways to protect privacy in EMR [4], which can include data protection protocols, hashing, data encryption, privacy anonymity processing, and access control.

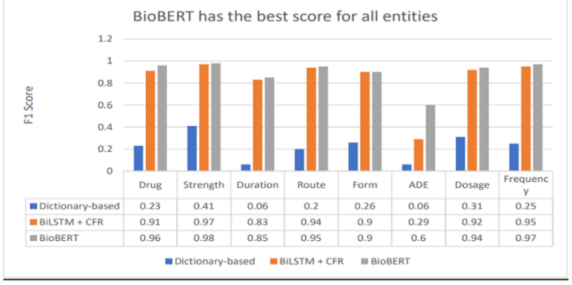


[Fig.1: Flowchart of EHR Data Collection & Assimilation]

III. RESULTS AND DISCUSSION

We use unstructured medical data in EHR to build models for biomarker extraction and standardization [1]. We demonstrate that we can apply deep learning NLP models using coevolutionary neural networks (CNN) or recurrent neural networks superior to classic machines to determine the status of patients' biomarkers accurately [5]. Having said that, the tremendous volume of unstructured texts in EHRs is one of the major barriers to artificial intelligence (AI) adaptation in medicine. We show here that a deep learning NLP algorithm is used to achieve high performance in EHRs with mixed languages and significant heterogeneity in the target parameters and values for extraction and standardization of biomarker values.

The Rule-Based model did not perform very well, but that was expected as it does not take context into account and has a very high false positive rate. For BiLSTM + CRF and BioBERT, the ADE-corpus dataset is integrated into our data which thereby improves the performance to a great extent. The F1 score for the ADE entity improved from 0.3403 to 0.8673 for the BioBERT model after adding a sample of the ADE Corpus.



	Drug	Strength	Duration	Route	Form	ADE	Dosage	Frequenc y
■ Dictionary-based	0.23	0.41	0.06	0.2	0.26	0.06	0.31	0.25
■ BiLSTM + CFR	0.91	0.97	0.83	0.94	0.9	0.29	0.92	0.95
■ BioBERT	0.96	0.98	0.85	0.95	0.9	0.6	0.94	0.97

[Fig.2: Performance of our Models Based on Different Attributes]

Although BiLSTM is pretty much similar to BioBert with respect to relatively all entities, there is one entity where the problem arises: its adverse drug event [3]. This is due to the lack of clinical trial data available to us. ADE is much more prevalent in the case of new drugs [6], so appropriate data must be made public in order for the models to work more efficiently as far as this entity goes [7].

Furthermore, we found that incorrect predictions of high confidence were primarily caused by human mistakes, and those with model mistakes were restricted to people with low trust [8]. Despite human errors in the training data set, the ability to achieve high precision highlights the stability of the method [9]. It should be pointed out that other errors in medical records and records, apart from manual annotation errors, are known to occur at error rates of up to 27%. The final models can be integrated into the overall pipeline to make it easier to automatically extract and normalize the biomarker values [10]. For subsequent manual validation, low-confidence forecasts can be marked [11].

IV. CONCLUSION

So, to conclude, we have developed and validated some clinical named-entity recognition models for free-text electronic health records. Information extracted from EMR text can be used to identify varied conditions and/or symptoms of patients with a high level of success rate. We also demonstrated that machine learning plays an essential role in developing a robust model applicable across different clinical domains.

The developed model too doesn't require an expensive infrastructure and can be used on standard machines with a decent CPU system. Further research is needed to improve the recognition of naturally underrepresented concepts. Future NLP studies should concentrate on the investigation of symptoms and symptom documentation in EHR free-text narratives. Efforts should be undertaken to examine patient characteristics and make symptom-related NLP algorithms or pipelines and vocabularies openly available.

DECLARATION STATEMENT

I must verify the accuracy of the following information as the article's author.

- **Conflicts of Interest/ Competing Interests:** Based on my understanding, this article has no conflicts of interest.
- **Funding Support:** This article has not been sponsored or funded by any organization or agency. The independence of this research is a crucial factor in affirming its impartiality, as it has been conducted without any external sway.
- **Ethical Approval and Consent to Participate:** The data provided in this article is exempt from the requirement for ethical approval or participant consent.
- **Data Access Statement and Material Availability:** The adequate resources of this article are publicly accessible.
- **Authors Contributions:** The authorship of this article is contributed solely.

REFERENCES

1. Baer B, Nguyen M, Woo EJ, Winiecki S, Scott J, Martin D, Botsis T, Ball R. Can Natural Language Processing Improve the Efficiency of Vaccine Adverse Event Report Review? Methods Inf Med. 2016;55(2):144-50. Epub 2015 Sep 23. PMID: 26394725. doi. https://doi.org/10.3414/ME14-01-0066
2. Ruud KL, Johnson MG, Liesinger JT, Grafft CA, Naessens JM. Automated detection of follow-up appointments using text mining of discharge records. Int J Qual Health Care. 2010 Jun;22(3):229-35. Epub 2010 Mar 27. PMID: 20348557. doi. https://doi.org/10.1093/intqhc/mzq012
3. Rochefort, C. & Verma, Aman & Eguale, T. & Buckeridge, David. (2015). O-037: Surveillance of adverse events in elderly patients: A study on the accuracy of applying natural language processing techniques to electronic health record data. European Geriatric Medicine. 6. S15. doi. https://doi.org/10.1016/S1878-7649(15)30050-4
4. Kalra, Dipak & Singleton, Peter & Milan, J & Mackay, J & Detmer, D & Rector, Alan & Ingram, David. (2005). Security and confidentiality approach for the Clinical E-Science Framework (CLEF). Methods of information in medicine. 44. 193-7. 10.1267/METH05020193. doi. https://doi.org/10.1055/s-0038-1633945
5. St-Maurice J, Kuo MH, Gooch P. A proof of concept for assessing emergency room use with primary care data and natural language processing. Methods Inf Med. 2013;52(1):33-42. Epub 2012 Dec 7. PMID: 23223678. doi. https://doi.org/10.3414/ME12-01-0012
6. Khare R, Li J, Lu Z. LabeledIn: cataloging labeled indications for human drugs. J Biomed Inform. 2014 Dec; 52:448-56. Epub 2014 Aug 23. PMID: 25220766; PMCID: PMC4260997. doi. https://doi.org/10.1016/j.jbi.2014.08.004
7. Shashi, Dr. M. (2022). Leveraging Blockchain-Based Electronic Health Record Systems in Healthcare 4.0. In International Journal of Innovative Technology and Exploring Engineering (Vol. 12, Issue 1, pp. 1–5). doi. https://doi.org/10.35940/ijitee.a9359.1212122
8. Patel, I., Jain, S., Vishwajeet, J. K., Aggarwal, V., & Mehra, P. (2021). Securing Electronic Healthcare Records in Web Applications. In International Journal of Engineering and Advanced Technology (Vol. 10, Issue 5, pp. 236–242). doi. https://doi.org/10.35940/ijeat.e2781.0610521
9. Hussain, Dr. M. K., Hussain, M. J., Bakri, M., Abdurraheem, T. M., & Al-Areefi, M. (2020). Big Data in Healthcare. In International Journal of Recent Technology and Engineering (IJRTE) (Vol. 8, Issue 6, pp. 2127–2131). doi. https://doi.org/10.35940/ijrte.f8100.038620
10. Khan, N. D., Younas, M., Khan, M. T., Duaa, & Zaman, A. (2021). The Role of Big Data Analytics in Healthcare. In International Journal of Soft Computing and Engineering (Vol. 11, Issue 1, pp. 1–7). doi. https://doi.org/10.35940/ijsce.a3523.0911121
11. Jeyaraj, B. Dr. P., & Narayanan AVSM, L. G. T. (2023). Role of Artificial Intelligence in Enhancing Healthcare Delivery. In International Journal of Innovative Science and Modern Engineering (Vol. 11, Issue 12, pp. 1–13). doi. https://doi.org/10.35940/ijisme.a1310.12111223

AUTHOR PROFILE

Aman Mukhopadhyay, the author is based out of West Bengal, India; and is currently a working professional proficient in software testing, cybersecurity, networking, IT governance just to name a few. He holds a B.Tech degree in Computer Science & Engineering with specialization in Bioinformatics from VIT Vellore, with numerous certifications and awards to his belt. Moreover, he has also made notable contributions in the field of image processing, his most recent work includes a research paper and the incorporation of a new technique that facilitates detection of arterial capillary vessels in retinal images.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of the Lattice Science Publication (LSP)/ journal and/ or the editor(s). The Lattice Science Publication (LSP)/ journal and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.