Artificial Intelligence Applications in Clinical Pathology: Refining Structured Data from Unstructured Medical Reports

Vayun Malik,†[a]

Mr. V. Malik
Department of Chemistry
E-mail: vayunmal@gmail.com

Abstract: In the current medical field, Unstructured data causes millions of pieces of information to be lost every year as translations make compounding errors. This is especially prevalent in medical reports as doctors are infamous for writing in indecipherable text for prescriptions. This has caused many to get the wrong medicine for their health problems. This issue can be solved through Natural Language Processing as I have done through my research. I created a short code that demonstrates how this can be done by taking unstructured data and converting it to structured data in the form of easy-to-read and understandable CSV files. This will help to remove some of the common errors that come from non-statisticians working with data. If the data is well-structured doctors that might not be as well versed in statistics as their peers will still be able to make sense of the data in easy-to-read CSV files. If these errors can be significantly reduced millions of lives will be improved as HCPs and PCPs will be able to give them the right medications they need to stay healthy, Additionally, with better and easier-to-understand information doctors can make better decisions during surgeries based on a patients health records if they are structured. This program is similar to other models such as SpaCy and NER which are the future of Natural Language Processing. They will lead the way to eliminating all unstructured data in healthcare and beyond.

1. Introduction

The medical field continues to expand as new technologies allow doctors to perform more complex and life-saving surgeries. However, with this new technology comes a problem for hospitals as it is impossible to store all of this technology in one hospital in the hopes that a doctor that specializes in this field can come in and use it. To solve this problem, hospitals must know what their up-and-coming doctors are going to specialize in so they can purchase the specific equipment for said doctors. I've done so by creating a Machine Learning model that can predict doctors' specialties when given certain information about them so that hospitals can best prepare when the time comes for the doctor to perform a specific surgery. Going into this project my main objective was to develop this model as best I can and push it the farthest that I can in terms of accuracy and reducing the number of predictors it needs while still being accurate. First, we must understand what unstructured data is. It is any type of data that is not in an easily digestible format. Unstructured data is everywhere in the modern world and over 80% of the data that is created every day is unstructured. (Lyng) This sheer volume of unstructured data makes it a large problem in today's technology-based society. It is very difficult for computers to efficiently sort through data and find what they are looking for. This is an especially big problem for the medical field as misinterpreting data can be the difference between a patient living or dying. This is especially apparent when it comes to prescriptions as doctors are infamous for writing prescriptions that are easy to misinterpret. Structured data is key to solving this issue as it allows information like these prescriptions to be understood and digested easily by humans and computers alike.

Vayun Malik was born in Cleveland, Ohio, in 2005. He is currently a senior at Thomas Jefferson High School for Science and Technology. He is interested in Computer Science and more specifically AI and plans to study it for his undergraduate and master's degrees. He is currently working on a website called Greeny Buddy that helps people understand their carbon footprint and gives them suggestions to reduce their carbon emissions.

2. Leading Natural Language Processing Models

This section discusses some of the most powerful Natural Language Processing(NLP) models in the field and their significant applications. Specifically, it focuses on SpaCy and NER as they are very powerful specifically in working with unstructured data that doctors often give out in the form of prescriptions and notes they take during checkups and examinations.

2.1 SpaCy Model and its Applications

In the last decade or so healthcare providers(HCPs) have transitioned to having most of their data inputted into computers. This should've put an end to unstructured data's prevalence in the medical field and the harm it has on it. However, HCPs failed to organize their means of entering data, and now each hospital system has its way of organizing its data. This makes it very difficult for HCPs to properly treat patients that are outside their network. Some of the most important of these elements are "problem lists, medications, and allergies" (Murray and Berberian). If a HCP can't tell what their patient is allergic to or what medications they've taken there is a significant risk of the patient being prescribed the wrong thing. If the data were all structured similarly then this wouldn't be a risk as HCPs would know where to look to find the information they need to treat their patients best regardless of if they've treated the patient before. One of the best ways to create this structured data is to use the SpaCy model which incorporates Natural Language Processing(NLP) to effectively process data.

SpaCy is a tool designed for Python that is very useful for NLP and was released in 2015. It is very intuitive to use and allows NLP to be done at a much larger scale than previous tools. It works in over 70 different languages and can detect nouns, verbs, adjectives, and other parts of speech in all of the said languages. It was rigorously tested for accuracy and was highly accurate in determining which part of speech was which. SpaCy is a very powerful tool and shows how powerful the field of NLP is and how far it can develop.

Some of the most common household items use NLP and models like SpaCy to do so. One example of this is Alexa which is depicted in the figure below. Alexa uses AI, NLP, and Machine Learning to understand what its users are saying and how to best fit their needs.

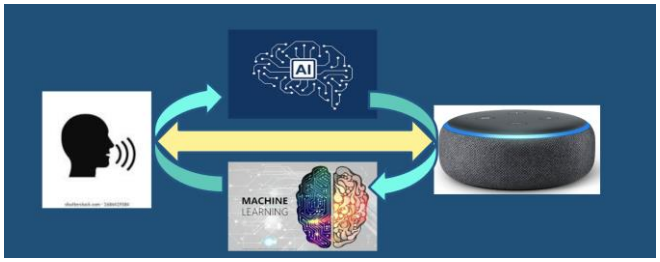

Figure 1. Schematic application of AI and NLP to the Alexa device

Alexa's AI and models like it go through a pipeline similar to the one below to convert unstructured data to structured data. They use NLP and AI to take unstructured data which is your voice in the case of Alexa and convert it into structured data which in the case of Alexa is its answer to your question which is normally based on sources it finds on the internet.

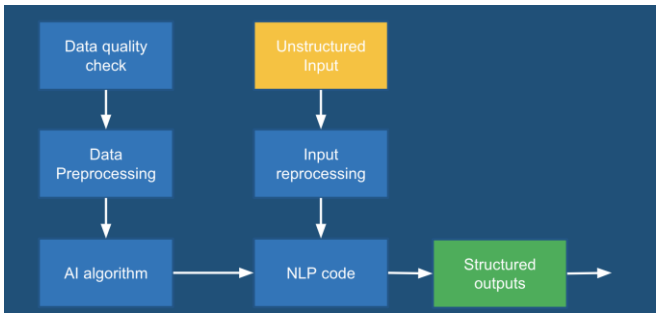

Figure 2. Pipeline of obtaining structured data and the validation process part of AI/ML architecture.

2.2 Named Entity Recognition

In the pipeline, the spot where the NLP code is the part that is the most versatile based on what type of NLP is meant to occur. One of the most important types of NLP is called Named Entity Recognition(NER) and it is one of the main things search engines and devices like Alexa use to give users the results they want. NER works by finding the most important words or phrases in an unstructured piece of data and searching its database for entities that match these words or phrases. Then it structures these entities and outputs them to the user and has completed its step of processing an unstructured language-based input into a structured language-based output.

NER is commonly used through the SpaCy model as it comes built into SpaCy. The model SpaCy uses is a "supervised deep learning model"(Yu) and it is a very effective implementation of NER. This model is normally trained with 10 epochs and after only the first 2 it becomes very accurate at recognizing the named entities. (Yu) This combination of NER and the SpaCy model is something the medical field absolutely needs in order to take a patient's unstructured symptoms and give a structured diagnosis through the help of a specialized Primary Care Provider(PCP). NER helps in this case by recognizing which symptoms are important and feeding them to a doctor in a structured and understandable format. Scheme 3 shows this in action below:

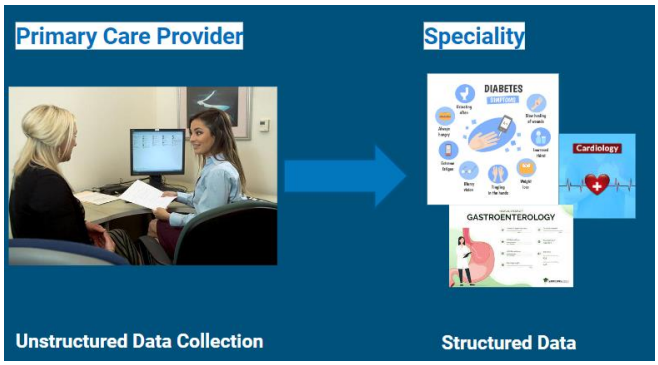

Figure 3. Spectrum of Unstructured data from PCPs to Speciality through NLP

3. Program Developed

This section discusses the program that was developed to convert unstructured data to structured data. This program takes an unstructured text file and uses NLP to create a structured CSV file that clearly represents the data that was put into it.

3.1 Code Algorithm

The applications of NLP are truly limitless as SpaCy and NER are just scratching the surface of what is to come for the field. NLP can be applied to help PCPs by using it to convert the unstructured data their patients give them to structured data that they can use to properly diagnose them. As shown below in Scheme 4 I used the following code algorithm to help alleviate this issue.

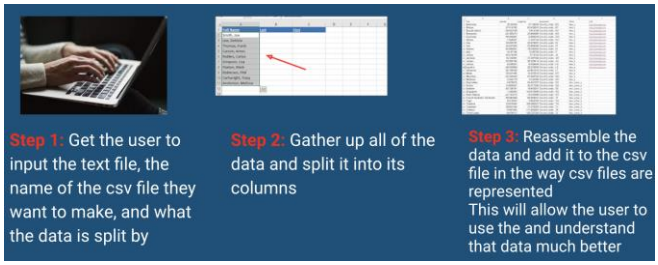

Figure 4. Algorithm development process and NLP Code generation

As seen in the figure above there are three main steps that my code does in order to transform the unstructured data a patient gives to structured data that a doctor can use and clearly understand. The first step is all about getting the necessary information from the user such as where the data is, where they want the structured data to be output to, and how the data is separated. After this step, my algorithm takes it from there as it splits the data based on the user's input from before and places it in lists with each list being a row of the eventual spreadsheet. Finally, my code puts all of these rows together and places them into a CSV file so that the user can see how the structured data looks and make their diagnosis based on the symptoms displayed.

Conclusion and Outlook

Unfortunately, unstructured data is very common in the modern world and it is a plague that allows misinformation to be spread and false diagnoses to be made. NLP is the key to fixing this issue and allowing structured data that is easy to understand and interpret to become more widespread in the medical field and

even beyond it. The power of tools like SpaCy, NER, and my own is essential to stopping unstructured data from ruining more lives. HCPs and PCPs need structured data as soon as possible and I hope that my research allows it to become more widespread in the medical community through clean and easy-to-read CSV instead of messy prescription notes and lists of symptoms that are easy to misinterpret.

Acknowledgments

The author thanks to Dr. Rajagopal Appavu for his mentorship and proofreading the manuscript.

Keywords: Natural Language Processing • Structured Data • Unstructured Data • Machine Learning • Artificial Intelligence

References

Blogger, G. (2011, March 31). The importance of structured data elements in EHRs. Computerworld. https://www.computerworld.com/article/2470987/the-importance-of-structured-data-elements-in-ehrs.html

Bonthu, H. (2021, August). How to Read and Write With CSV Files in Python? Analytics Vidha. Retrieved August 1, 2023, from https://www.analyticsvidhya.com/blog/2021/08/python-tutorial-working-with-csv-file-for-data-science/

Grossman, J., & Pedahzur, A. (2020). Political Science and Big Data: Structured Data, Unstructured Data, and How to Use Them. Political Science Quarterly, 135(2), 225-257. https://doi.org/10.1002/polq.13032

Honnibal, M. (2021). spaCy (Version 3.6) [Computer software]. Explosion. https://spacy.io/

Huang, Y. (2022, April 12). Clinical Named Entity Recognition Using spaCy. Towards Data Science. Retrieved August 1, 2023, from https://towardsdatascience.com/clinical-named-entity-recognition-using-spacy-5ae9c002e86f

IBM Cloud Education. (2021, June 29). Structured vs. Unstructured Data: What's the Difference? https://www.ibm.com/blog/structured-vs-unstructured-data/

Inmon, W., & Linstedt, D. (2015). Unstructured Data. Data Architecture: A Primer for the Data Scientist, 63-70. https://doi.org/10.1016/B978-0-12-802044-9.00011-8

Lyng, G. (2021, October 4). 4 Risks of Storing Large Amounts of Unstructured Data. Dataversity. Retrieved August 1, 2023, from https://www.dataversity.net/4-risks-of-storing-large-amounts-of-unstructured-data/

Meaney, A. (2021, May 11). Working with CSV files. https://www.fundrecs.com/blog/working-with-csv-files

Moskalev, I. V., Krotova, O. S., Khvorova, L. A., & Bobkova, D. G. (2020). Extraction of structured data from unstructured medical records using text data mining technologies: process automation. Journal of Physics: Conference Series, 1615(1)https://doi.org/10.1088/1742-6596/1615/1/012031

Pallamala, R. K., & Rodrigues, P. (2022). An Investigative Testing of Structured and Unstructured Data Formats in Big Data Application Using Apache Spark. Wireless Personal Communications, 122(1), 603-620. https://doi.org/10.1007/s11277-021-08915-0

Patel, D. (Director). (2022). Named Entity Recognition (NER): NLP Tutorial For Beginners - S1 E12 [Video]. Codebasics. https://www.youtube.com/watch?v=2XUhKpH0p4M

Schafer, C. (Director). (2017). Python Tutorial: CSV Module - How to Read, Parse, and Write CSV Files [Video]. Corey Schafer. https://www.youtube.com/watch?v=q5uM4VKywbA

What is Named Entity Recognition (NER)? (n.d.). Techslang. Retrieved August 1, 2023, from https://www.techslang.com/definition/what-is-named-entity-recognition-ner/#:~:text=Named%20Entity%20Recognition%20(NER)%20is,Entity%20Extraction

Aparna, C., Kaviya, G, Deepa, M. (2023) Analyzing the Function and Structure of Components within the mRNA COVID-19 Vaccination for Determining the Cause of Anaphylaxis Reactions within Vaccine Recipients 2023 Feb 7; 1(1):0001-0005

What is natural language processing? (2023, January 6). IBM. Retrieved August 1, 2023, from https://www.ibm.com/topics/natural-language-processing