Document ID CGEP010121

Redefining Clinical Trial Decision-Making: Leveraging NLP for extracting insights from unstructured data

Saurabh Das- Head, TCS ADD™ Research and Innovations; Sushil Kumar Singh- Associate Consultant, TCS ADD™ Analytics and Insights; Niketan Panchal- Researcher, TCS ADD™ Analytics and Insights; Rajasekhar Gadde- Researcher, TCS ADD™ Analytics and Insights; Rohit Kadam- Researcher, TCS ADD™ Analytics and Insights

Abstract

In the entire process of clinical trial execution, numerous feedbacks, comments, reviews, physician notes and many other textual data is captured. These free texts even though they are captured across multiple systems. It so happens that these free texts are rarely picked up for processing of signals and subsequent management, generating meaningful insights, inferencing, managing risks, compliance etc.

Leveraging next-gen technologies in data ingestion and extraction, can enable systems to seamlessly extract and analyse such free texts. Natural Language Processing (NLP) has progressed to the point where it cannot just classify the text into categories but also facilitate in highlighting the essence of the message or intent conveyed via given text. This information then can be used in tandem with statistical and deep learning models for further classification and to generate actionable insights and oversight. The gestation period of clinical trials is very long and the data that is generated through the entire cycle is continuous and can be obtained during various phases of the trial. This data is progressive over the entire duration of a clinical trial and hence, utilizing this data can lead to some critical insights pertaining to certain trends and patterns mapped to the trial with respect to risks, compliance or safety aspects. This, in turn, can enable quicker and data-driven decision making.

Free texts can be categorized, ranked or prioritized, grouped or clustered into similar texts and extracted, all by leveraging certain NLP and deep learning algorithms

Natural Language Processing (NLP)

Natural Language processing (NLP) is an area of computer science, which deals with the problem statement of making computer systems understand natural language text and speech to perform certain tasks [1]. Over time, NLP techniques have evolved from extraction of texts to interpretation of it and eventually generation of new texts all by itself. The NLP tasks can be broadly categorized into text classification, Named Entity Recognition (NER), question- answering, summarization and Text Generation (*simply known as Natural Language Generation*).

Throughout the course of this article, we will focus on 'Text Classification'. The process of understanding the text and then categorizing the given text into predefined buckets is known as text classification. For instance, we have a dataset of physician notes; our job is to categorize notes into categories 'having adverse event' and 'not having adverse events'. For this task we manually classify each physician note into either of the category. This annotated dataset comprising of notes and their corresponding labels is capable of performing text classification when fed to a machine learning model. It will learn the pattern from the natural language text, hence, when a new physician note is given as an input to the model, it can quite easily predict if the note has information about adverse event or not, all by itself.

Deep Learning based NLP techniques further enable a machine to understand the context of any sentences/phrases and then categorize the sentences accordingly. That being said, sentences that

mean the same can be grouped together and formed clusters, hence, paving the way fora better understanding and visualization of the data.

Life Sciences

Life Sciences is a branch of medical domain which predominantly encompasses the field of drug discovery to drug development, conducting clinical trial of the developed drug and monitoring the drug safety and efficacy till the drug reaches the post marketing surveillance phase.

Drug discovery and development involves a lot of genomics and research work and when a drug molecule is finalized for a given therapeutic area, the clinical trial process begins. The trial process involves a molecule to undergo multiple phases before being approved by the regulatory bodies for marketing. These trials are conducted to thoroughly test the drug molecule for its efficacy, safety, dosages etc and involves a high level of complexity and numerous stakeholders. A huge amount of data is generated in this process which needs to be analysed for generation of meaningful insights and oversight. The data is in various formats such as text, images, graphs and has multiple systems and pipelines in place in order to ingest, store and perform actions over them. However, a point to note is that the majority of this data is in text form, vis-à-vis, structured text (i.e. in form fields) and as unstructured text. (i.e. descriptive text)

Unlocking Insights from unstructured data in RBQM

Risk-Based Quality Management (RBQM) is a system for managing quality throughout a clinical trial in tandem with risk-based monitoring. It is a data-driven strategy that has evolved substantially over the past few years, as an extension to the original principles underpinning risk-based monitoring (RBM) [3]. RBQM focuses on all the things that matter in a clinical trial; starting from planning till execution [4].
RBQM is a combination of tools; it provides a centralized monitoring platform which helps in central data review, risk assessment, data quality oversight, evaluating KRIs and issue & action tracking management modules.[3]
The RBQM system receives input from multiple diverse source systems, analyses those inputs, and monitors and forecasts risk mapped to a specific site. As part of the form-based data collection (Electronic or Paper Based) in the complete process of clinical trial execution, the obtained data may be broadly classified as structured, i.e. fixed fields and unstructured, such as feedbacks, comments, reviews, nurse's notes etc. Traditionally, structured text is utilized for reporting, analysis, and gaining insights. Unstructured sections and fields in documents or data sources are hardly ever processed as a result of the intricacies involved.

With the current RBQM systems, there is a limitation of handling unstructured data which involves large volumes of data available in siloes and are a source of hidden risks. Unseen signals are not picked up by the traditional RBQM systems as they are expensive & resource intensive. Identifying these risks will require cognitive intelligence or intensive human efforts.

To the best of our knowledge, no previous studies describe the analysis of hidden risk and signals directly from free text. It is important to emphasize that nowadays a lot of the patient information in the EHR's, monitoring visit reports, query management systems, MV issues, uncoded AE/CM/MED terms, central monitor's Response to signals, safety narratives all are in free text.

NLP techniques can be used in RBQM to parse through unstructured data and analyze them. Along with AI/ML this can also be used for predictive analysis. In the figure below, we explain the use of NLP and AI in RBQM. From various data sources input that obtained is passed through the NLP extraction engine. This engine extracts the textual data from documents and creates an

intermediate spreadsheet-like data format of the text data. Considering Medical Visit Reports (MVRs) as the input, the NLP extraction engine extracts all the comments from the MVR and prepares a structured data format. This data comprising of MVR comments is then given as input to the NLP prediction engine.

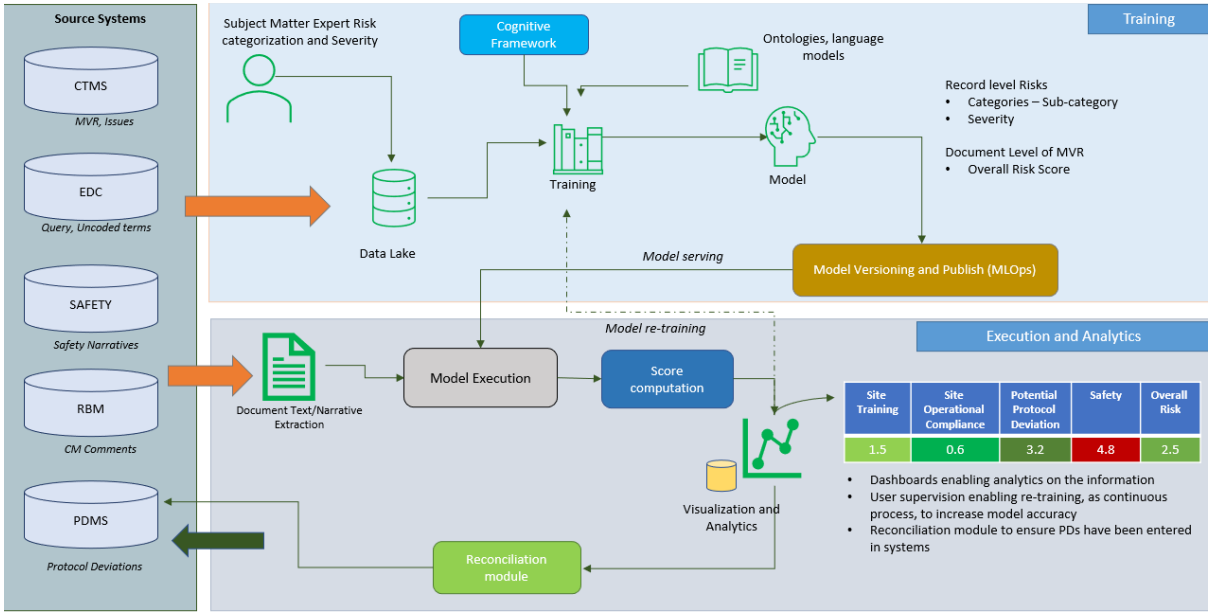

Figure 1 High Level Architecture of the system

NLP prediction engine is a deep learning model trained to perform the text classification task. This model has previously been trained on MVR comments, their corresponding protocol deviation (PD) category and severity score. It means that this model is capable of predicting the PD categories with their severity when an MVR comment is provided as an input. As depicted in figure 2, a statistical model for weighted factored computation can be used to derive the overall site risk score

Medical Visit Report Free Text	Site Operational Compliance	Site Training	Safety	Potential Protocol Deviation	Overall Site Risk Contribution
The person took a Conmed before the visit which makes him non-complaint for the trial	**20%**	**60%**	**10%**	**85%**	**4.3**
Eye care specialist to complete the worksheet fully as there are few sections in the sheet not being completed	**60%**	**80%**	**10%**	**5%**	**2.8**

Figure 2. Classification of Risk into Categories of Protocol Deviation

These severity scores and site score can be monitored over the course of a clinical trial. With multiple documents being uploaded and analysed over time, we will have a pattern/trend line generated for the site, depicting its risk score. This can be further used to check the overall risk to the site and predict which are the major categories contributing to the risks in the future and proactively work towards eliminating them.

Future Scope

Text present in the data is a valuable source of information for predictive model development and should not be neglected at any cost. A future outlook on explainability and external validation of the developed models will lay the foundation for robust and trustworthy predictive models in clinical

practice. Learning and deriving insights from historical data can help facilitate in data-driven decision making in multiple facets of clinical trials ranging from study setup, risk assessment and signals management.

With the advancements in natural language techniques and the recent developments in Large Language Models (LLMs), they can be used to solve this use case. LLM fine-tuned on a dataset, can act as a building point for multiple point solutions. The LLMs in instruction following mode can be used to perform certain tasks over the MVR documents like extracting specific information and insights and then creating a summary out of the given data. In chat mode, the LLMs can be used as an assistant to the human user which will try to answer the questions based on the input document, hence, providing a better explainability as to why this is given as an answer to a particular question. Certainly, LLMs open up a wide range of possibilities which can be explored and leveraged to solve business use cases.

References:

1.Chowdhury, G. (2003) Natural language processing. Annual Review of Information Science and Technology, 37. pp. 51-89. ISSN 0066-4200
2.https://www.quanticate.com/risk-based-monitoring#:~:text=RBM%20means%20that%20the%20volume,assessing%2C%20monitoring%20and%20mitigating%20risks.
3.https://acrpnet.org/2020/08/10/risk-based-clinical-trial-management-harnessing-the-transformation-of-rbm-to-rbqm/
4.https://globalforum.diaglobal.org/issue/september-2020/looking-to-the-future-with-rbqm/