The Challenges of Statistical Analysis and Unstructured Data.

***Abstract*— Unstructured data is now accounting for up to 80% of all data held by organisations. Currently, processing this data is time-consuming and complex, as traditional analysis methods are designed to be deployed against structured data. However, unstructured data hold the potential to answer many questions. In this paper the use of Machine Learning, including Deep Learning, and NLP, focusing on word embeddings and the use of word2vec and BERT with unstructured datasets is reviewed, with focus on medical and financial unstructured data. The paper summarises the state of the art and outlines the open research questions of privacy, quality and availability of datasets and generalisability of study outcomes that face the field, highlighting where they exist, some solutions to these questions that have been proposed in the literature.**

***Index Terms*— Machine Learning, Statistical Analysis, Unstructured Data**

I. Introduction

Unstructured data can be considered the opposite of structured data. Structured data is that which has a well-defined structure and is stored in a manner that adheres to this documented and defined schema, for example a relational database. Sitting in between the two is semi-structured data, where the data maybe incomplete, inconsistently structured or have a structure that changes quickly. Additionally, the data itself may change quickly or unexpectedly within this non-fixed schema. So therefore, that leaves unstructured data, which is data that does not fit into either of these descriptions [1]. Unstructured data has no fixed form, schema, or structure and because of this is not searchable, sortable, easily visualised or analysed. Common examples of this type are emails, images, ad-hoc spreadsheets, and emails. From this non-exhaustive list of examples, it is clear to see that a modern organisation possesses and generates a huge amount of this non-structured data. Indeed, studies put the proportion of data that is unstructured in the region of 80% [2].

This paper takes the use of statistical techniques when working with unstructured data in text form as its applied use case in order to define the scope of the paper and provide contextual basis for discussion. This unstructured text category has received a substantial amount of interest [3], ranging from the processing of invoices and receipts to the analysis of medical notes and records, in both hand-written and word-processed forms. It is these two domains that this paper will further restrict itself to, for the purposes of discussion and illustration, drawing on the scholarly interest in this particular niche of the topic to follow lines of enquiry and identify state of the art and open research questions. Furthermore, this paper reviews the use of statistical methods to analyse unstructured data, specifically the use of statistical models for the analysis of text, including Natural Language Processing (NLP), Machine Learning (ML) models and a progression of these, Deep Learning and the challenges that still face the effective use of these tools, technologies, and methods for this problem.

The remainder of this paper is organised in the following sections. Section 2, Technologies which discusses the current state of the art and the methodologies, techniques and technologies used to work with unstructured data. Section 3, Open Research Issues, this section seeks to outline current frontiers of research related to the processing of unstructured data. This is followed by Section 4, Conclusion which provides a summary of the issues discussed.

II. Technologies

This section reviews the technologies employed when using statistical methods to analyse unstructured data.

A. NLP

Natural language processing (NLP) is the field that has developed around giving the computer the ability to understand text, and importantly the messages and meaning it is conveying. NLP is useful in the automated extraction of key features. In this chosen area, these are the specific fields of interest to the problem, be them the name and address of the company sending the invoice, the amount the invoice if for, or the name of the drug prescribed, and the quantity. In particular, Named Entity Recognition (NER) in the field of medical notes and invoices, can be especially tricky, as due to their unstructured nature the task is harder than a standard NLP task that is performed upon structured data. Ref. [2] explains that understanding the document layout is of crucial importance in understanding the document at all, this is made all the more complex a task when it is considered that many of these documents are either handwritten entirely, in the case of medical notes, and or, often annotated as is the case with invoicing in a manner that is seemingly random. To address this need for improving capability in the realm of text extraction and analysis, which is recognised by the International Conference on Document Analysis and Recognition (ICDAR) who arrange competitions in a bid to advance the field. They all provide dataset for training, notably for this context a dataset comprised of receipts to support this aim.

B. Word Embedding

Word embedding is a technique used within NLP that assigns words or phrases to a vector of corresponding values. This can be complex and time consuming when performed on medical notes, due to the length of the notes and the use of multiple timestamps. Having these multiple time stamps increases the possibility of words and phrases having multiple meanings within the text, depending upon the time that they appeared. This increases complexity as the time also needs to be captured

within the embedding. However, the advantage of using embedding techniques is their unsupervised nature – that is they do not require labelled data up front. This allows embeddings to be learned, in an unsupervised way, from unstructured data and then be applied to supervised learning at the next, or later stage in a pipeline [4].

There are several established word embedding models, Word2Vec, a popular, early model in the field, uses either a continuous bag of words model, or a skip-gram model. A neural network is then trained to, in the case of continuous bag of words predict an unknown word based upon those known words that are close to it both before and after the target, or in the case of skip-gram, predict the probability of words in the same context, or a defined range, being close by to the given word.

This has been further expanded upon by the BERT (Bidirectional Encoded Representations from Transformers) model. BERT is a model that has been pre-trained on a huge dataset that is a labelled corpus and is capable of performing NER tasks with a contextual understanding of the data it is working upon. This model uses information gained from the words before and after the target and trains a deep neural network that can predict the masked word. The whole of network is kept post-training which can has to be fine tunned later to fit the new dataset. Studies have shown that with some fine tuning the performance can be further increased to gain 5% in NER tasks.

C. Machine Learning

Machine learning (ML) has been deployed in studies [2], [5] aimed at classifying text contained within these selected types of documents. A range of methods and models have been reported upon, including Support Vector Machines (SVM), Naïve Bayes and Hidden Markov Model (HMM) to name three. This isn't a simple solution, however [2] goes on to report that there are several challenges that face the deployment of these models, singling out Naïve Bayes and SVM in particular. These challenges include the fact that the models, as they stand, are not capable of completing the whole of the cycle from end to end, including extraction, rather they are suited to the classification and field recognition tasks. ML models require data for training, and often lots of it. This data also needs to be labelled in most cases, and this is problematic when it comes to unstructured data as this data is time consuming to label, and when done manually negates the advantage of ML.

D. Deep Learning

Deep Learning (DL) is a potential solution to some of the problems previously discussed above. Named Entity Recognition is an area where DL and ML provide significant assistance. NER is the pivotal task within text extraction of identifying the entities or significance or interest, for example a patient, drug, company, invoice number, address and so on. This is a crucial step in the process as analysing texts is significantly simpler with these features identified. DL is well suited to automating feature extraction and identification via the use of pre-trained neural networks (NN). Where the size of the available dataset is large enough, it is possible to use NNs to learn these features from unlabelled data, both Recurrent Neural Networks (RNN) and Convolutional Neural Networks (CNN) are deployed in exactly this scenario to good effect [2].

Each of these models have their own distinct advantages. CNN has been used for feature extraction in several studies, as discussed in [2]. However, research has focused on the use of RNN which has an architecture that is more complex than CNN, but this complexity is due to the fact that the output from previous model runs is fed back into the network. This allows for the tokens that represent the text to be known/stored by the model and therefore the sequence that may contain sematic information is captured.

III. Open Research Issues

In this section the issues that remain open and that are actively under research and further investigation will be discussed, through the lens of the technologies discussed above. These will be broadly grouped into the following issues to frame the discussion further, Privacy, dataset availability and data quality.

A. Privacy

Medical and Financial datasets in particular are subject to strict privacy regulations including General Data Protection Regulations (GDPR) in Europe and, for medical data, Health Insurance Portability and Accountability Act (HIPAA) in the USA [3]. These restrictions present a range of challenges for researchers, analysists and data scientists wishing to work on problems in these domains, and it is acknowledged that automatically deidentifying using NLP methods would increase the availability of data that could be used for research [6].

To address these challenges work has been undertaken and [3] suggests that this is taking two main directions. One is the use of pseudonymisation and deidentification to ensure that the data does not contain information that can lead to the identification of a patent in the records. The attraction of this direction of travel and technique is that there are already processes and techniques that can achieve this goal. The second direction is to take the methods and techniques that are to be employed and make adaptations and changes to them so that they are able to function in a way that preserves the privacy of the patients in the data. This isn't without its own additional challenges however, as these changes are usually accompanied by increases in cost, complexity, and reliability of the adapted algorithms. Both of these open issues are discussed in further detail in the following paragraphs.

B. Deidentification and Pseudonymisation

Simply put, deidentification is the removal of identifying information from a dataset, in the context of medical notes these identifiers would usually be patient identifiers, patient names and other personal details. Pseudonymisation is where theses values, instead of being removed from the dataset entirely are replaced with new values, known as surrogate values. This technique has the advantage of preserving the integrity of the data but still removing the link to an actual person. These surrogate values are used consistently throughout the process so that the new surrogate value(s) reference the same original individual in the dataset.

While a valid technique, the open questions that surround this issue are not focused upon the validity of using deidentification or pseudonymisation, but on the application of these techniques. Specifically, the questions and areas in need of further research are linked to how the values that need removing, or anonymising are identified in a dataset to start with. This issue is complex in large part as a direct result of the complexity of the datasets, and in particular the unstructured nature of them.

Ref. [3] identifies two studies that have demonstrated the ability to perform these tasks using computation methods to a high degree of accuracy, reporting F1-scores of 0.9878 and 0.9584 for identifying personal identifiers within the clinical notes dataset using Conditional Random Fields and regular expressions against a dataset in Chinese and using deep learning against an English dataset respectively. While encouraging, and certainly supportive of the techniques application to solving this issue, the question remains open as it is still unclear if these are specific results, or if the techniques open the door to generalisable methods that can be applied to datasets comprising of different languages, and crucially for this medical note domain, the notes created and maintained by different organisations.

C. Availability of datasets

Having access to high quality datasets from which models can be trained is a pressing problem. This is especially true, but not limited to the medical and finance areas, where there are additional barriers, such as the ones previously discussed related to confidentiality.

The access to gold-standard datasets, and in particular labelled data is a challenge [3], both in terms of availability and also in terms of dataset size. This is in part, in all domains not just medicine and finance although the impact is acute there, that the data within these datasets is very specific to the organisation, and or, of commercial value or subject to other constraints that make the organisations that poses it to be unwilling to make the data publicly available. Consider a dataset of labelled invoices, this dataset would contain many sensitive items, name and address of customers, product pricing, and other commercially sensitive information that make it unsuitable to public consumption.

To produce a high-quality dataset takes time and domain specific and specialised knowledge and there is work being undertaken to address this, often leveraging supervised and semi-supervised machine learning techniques [3]. An example of this is the use of clustering to label data initially using the centre of the clusters before feeding these labels into a supervised methodology.

When considering data that's is required for NLP problems there are defined methods already available to address this challenge. For example, using natural language generation tools or other artificial intelligence techniques to produce text that is synthetic but available for use in training NLP models. This technique has its drawbacks however, as there are open questions around the quality of the data these techniques produce [3], which of course, lead to question the models produces by using this data for training.

Extraction of data even when the raw data is possessed adds to this challenge. For example, [2] suggests that when it comes to using invoices, treating them as any other NLP task isn't possible given the variety that exists within even a small number of invoices. This variation in layout, text size, number of pages and other structural differences make the processing of these unstructured documents complex. One approach to overcome this is to treat the task of extracting the key information or features as an images processing task, rather than a text parsing one. However, they also note that this can introduce additional problems, as any sematic meaning is lost in the process. Ref. [7] is more optimistic, suggesting that the challenge posed in being able to create structured representations of unstructured data is one that has attracted significant amounts of attention and effort from the medical informatics community focused upon making this unstructured data more understandable and useable in models. They go on to highlight that the methodology of choice to add structure to this data has been the use of embedding, an area previously discussed.

Ref. [6] suggests that to resolve the issue of data availability entirely new methodologies are required and present one such novel method to anonymisation via obfuscation. They do however also accept that their technique would not be appropriate in all situations highlighting one such case, the need to preserve human readability for interaction with the dataset. While suggesting possible next steps to advance this technique further it demonstrates that this is very much a live and open, ongoing area of research.

D. Quality

This availability issue extends to gaining access to data stored in central warehouses, leading to studies and models being completed using only local data from a single institution. This raises quality issues, as the data may not be representative and include biases or other issues meaning that the results are not generalisable. For this reason, aggregation it has been suggested that while aggregating data may not be possible, as discussed above, aggregation of the resulting models from a range of institutions and therefore datasets maybe able to proxy global datasets [8]. This approach, similar to federated learning, allows for models to be trained individually, within institutions, and following their data protection procedures before the results, which are free of patient data can be shared with others. This is not without its drawbacks though, using different embeddings results in models that are harder to analyse as even similar, or possibly identical events result in different embeddings. Encouragingly, [8] publish a proof of concept for the harmonisation of the disparate models arising from different hospitals leading to better predictions. This has additional significance as along with improving data size and quality it also addresses the previously discussed privacy issue to a point that they are prepared to guarantee patient confidentiality. By their own admission, however, this requires further work, not least as they restricted the study to a single type of medical database, but they also acknowledge the end goal should surely be harmonisation to a global model, and this is not yet possible.

E. Generalisability

Generalisability is the notion that a model is transferred from one dataset, that it has been trained with and deployed using another, say for example another institution entirely and is still able to produce acceptable level s results. As has been touched upon in earlier paragraphs, the availability of findings that are of use outside the initial tightly defined field of study are few and far between. In the best cases models are produced that require only some fine tuning. Being able to create models, or produce frameworks that are generalisable across domains, and even in most cases across institutions is still very much an open research issue. However, [3] are optimistic about the future holding solutions in this area and are confident that tools developed using the methods discussed will be available and in use in hospitals and GP surgeries.

IV. Conclusion

It is clear that unstructured data could add tremendous benefit to healthcare by improving the use of data to both support patient care and medical research. There are also clear productivity gains to be realised in the field of finance where automation of tasks can be implemented. To realise this potential however, there are several barriers that need to be overcome. Some are practical, such as the implementation of known anonymisation techniques to preserve patient and client confidentiality, and others more technical and cantered around the ability to apply statistical models and techniques the multitudinous data that both industries possess but is in an unstructured form.

This report has summarised the key areas of privacy, availability, quality, and generalisability where there are open research questions. Here while the questions remain open, there is clear indications that these problems are currently being addressed, and in key, albeit narrowly focused areas, demonstrating that progress can be made.

References

[1] A. C. Eberendu, “Unstructured Data: an overview of the data of Big Data,” vol. 38, no. 1, pp. 46–50, Aug. 2016, doi: 10.14445/22312803/IJCTT-V38P109.

[2] D. Baviskar, S. Ahirrao, and K. Kotecha, “Multi-Layout Unstructured Invoice Documents Dataset: A Dataset for Template-Free Invoice Processing and Its Evaluation Using AI Approaches,” vol. 9, pp. 101494–101512, 2021, doi: 10.1109/ACCESS.2021.3096739. [Online]. Available: https://ieeexplore.ieee.org/document/9481217

[3] M. Tayefi et al., “Challenges and opportunities beyond structured data in analysis of electronic health records,” vol. 13, no. 6. Wiley, 14-Feb-2021.

[4] N. Oubenali, S. Messaoud, A. Filiot, A. Lamer, and P. Andrey, “Visualization of medical concepts represented using word embeddings: a scoping review,” vol. 22, no. 1. Springer Science and Business Media LLC, 29-Mar-2022.

[5] D. Zhang, C. Yin, J. Zeng, X. Yuan, and P. Zhang, “Combining structured and unstructured data for predictive models: a deep learning approach,” vol. 20, no. 1. Springer Science and Business Media LLC, 29-Oct-2020.

[6] M. Abdalla, M. Abdalla, F. Rudzicz, and G. Hirst, “Using word embeddings to improve the privacy of clinical notes,” vol. 27, no. 6, pp. 901–907, Jun. 2020, doi: 10.1093/jamia/ocaa038. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/32388549

[7] N. Oubenali, S. Messaoud, A. Filiot, A. Lamer, and P. Andrey, “Visualization of medical concepts represented using word embeddings: a scoping review,” vol. 22, no. 1. Springer Science and Business Media LLC, 29-Mar-2022.

[8] Y. Huang, J. Lee, S. Wang, J. Sun, H. Liu, and X. Jiang, “Privacy-Preserving Predictive Modeling: Harmonization of Contextual Embeddings From Different Sources,” vol. 6, no. 2. JMIR Publications Inc., 16-May-2018.