Quality Evaluation of a Documental Big Data Tool

[Authors information removed]

Keywords: Text Analytics, Big Data Analytics, Enterprise Content Management, Document Management.

Abstract: This paper presents the analysis of quality regarding a textual Big Data Analytics approach developed for documents. In the context of a project dealing with extraction of knowledge from document and process data in a Big Data environment, the paper focuses on performance and data confidence indexes to evaluate the quality of the underlying process. It shortly discusses some experimental results.

1. INTRODUCTION

In recent years, Big Data Analytics (BDA) has spread rapidly, proposing tools, techniques and technologies that allow extracting enterprise knowledge useful to companies and society [Wang 2018, Nicola 2014]. In this framework, this paper describes a solution for *document processing automation* based on *Enterprise Content Management (ECM)*, and focuses on evaluation of *quality indexes*. It presents an experimental BDA approach in a real setting, given by the studies performed in the "Big Data Project" concerning an innovative solution to Big Data. This project develops a solution shared among three companies, structured in a temporary agreement for competitiveness.
The focus here is on incorporating features of Big Data and process quality evaluation, in terms of both data quality parameters and process quality, such as efficiency and effectiveness, resulting in evaluation of service performance and of textual data quality, referred to the linguistic coordinate [Batini 2016].

"Quality" for the three companies applies to business analytics and data mining on company data sources, prediction on business processes, cross-selling analysis, and log analysis for automatic improvement of user experience and for document contents analysis. In particular, considering the *documents*, the project provides advanced analytics and Machine Learning (ML) on the document content and process data for optimization.

This paper points out the quality assessment and performance monitoring for textual BDA by illustrating the model, the system architecture, and the experimental results achieved in a cooperation between Politecnico di Milano and Microdata Group[1].

The paper is organized as follows. Section 2 presents the approach to Quality Evaluation, giving the research questions of the paper. Section 3 describes the system architecture and technologies. Section 4 focuses on process quality and data quality evaluation, giving the basic experimental results. Section 5 concludes the paper.

2 QUALITY EVALUATION: THE APPROACH

In building a BDA system, core data assets present many challenges, such as data quality, data integration and data security. In particular, data quality problems add complexity to the use of Big Data, with several general data quality challenges.
First, data are noisy, erroneous, or missing. For example, jargons, misspelled words, and incorrect grammars pose significant technical challenges for linguistic analysis. Moreover, data captured by mobile and wearable devices and sensors can be noisy. Data cleaning pre-processing occurs prior to our analytics system.

[1] https://www.microdatagroup.it/

Second, as data are growing exponentially, it becomes increasingly difficult for companies to ensure that their sources of data and information are trustworthy. Veracity of big data, which is an issue of data validity, is a bigger challenge than volume, velocity, and variety in BDA. It is estimated that approximately 20–25% of online consumer reviews texts are fake [Quiao 2017].
Data cleaning, filtering, and selection to detect and remove noise and abnormality from data automatically becomes essential. This activity involves statistical and analytical processes. Developing effective ways to detect and remove unauthentic data becomes critical to ensuring adequate trustworthiness of data. There are many partially or even unlabelled data.
Traditional supervised ML techniques require access to many labelled training samples. In some real-world Big Data projects (e.g., financial fraud detection), partially labelled data are considered as a type of noise/inaccuracy. Poorly labelled data cause problems when using ML algorithms for model building. Therefore, research involves semi-supervised machine learning.

2.1 Research Questions

Our research questions focus on the development of uniform data quality standards and metrics for BDA that address various data quality dimensions (e.g., accuracy, accessibility, credibility, consistency, completeness, integrity, auditability, interpretability, and timeliness).
The considered Project SIBDA: "Sistema Innovativo Big Data Analytics", aim was to demonstrate the potential of Big Data tools and methods for “Smart Cities” using data and sensor technologies, in combination with BDA, to deliver services in an *efficient* way. In particular, due to the growing relevance of *unstructured information* for enterprise business processes, Big Data are growing in mass and relevance for Smart Cities development and management. Smart Cities were the objectives of our “Big Data Project”. The research aims regard the creation of platform models for the collection and management of Open Data (Smart Platform), which should be replicable, so leading to standard solutions suitable for urban contexts of medium/small size. We refer the interested reader to [Fugini 2018, Fugini 2019] for details about the overall system in the Project.

In thus paper, our research questions focus on how to set up an evaluation system for the analysis of *performance* and *quality aspects* of the *BDA system*.

To this aim, the most significant techniques and the technologies considered in SIBDA are in the three following areas.
1. *Data ingestion*. Among the scenarios that characterize the stage of acquisition and initial processing of Big Data, one of the most relevant concerns data coming from IoT, with the enabling middleware and event processing techniques that support an effective integration [Marjani 2017]. For text analytics, the research questions regard ECM, namely how advanced content management applications are characterized by the growing significance of information extraction in the enterprise environment, and how the diffusion of Big Data storage tools applies to document-oriented databases.
2. *Data storage*. For BDA, we identify two research questions: i) how to store large volumes of data while maintaining high performance; ii) how to archive unstructured or variable data. Two other issues related to storage techniques also play an important role, namely: i) how to ensure the security of archived data; ii) how to limit the complexity and costs of the hardware infrastructure when scaling up volumes. Different storage models have been proposed in the presence of large volumes and yet adequate performance, particularly in response times. These new types of databases include NoSQL databases and NewSQL databases [Moniruzzaman 2013]. An interesting model is proposed for document-oriented databases. Values consisting of semi-structured data are represented in a standard format (e.g., XML, JSON - JavaScript Object Notation - or BSON (Binary JSON), and are organized as attributes or name-value pairs, where one column can contain hundreds of attributes. The most common examples are CouchDB (JSON) and MongoDB (BSON). This way of organizing information is suitable for managing textual content.
3. *Data analysis*. Analytics applications are the core of the Big Data phenomenon, as they are in charge of extracting significant values and knowledge from data acquired and archived with the techniques above. This result can be achieved either through Business Intelligence techniques, or through more exploratory techniques, which can be defined as Advanced Analytics [Chawda 2016]. The generated knowledge has to be available to users and shared effectively with all the actors of the business processes that can benefit from it. In this area, we

mention the techniques of Big Data Intelligence, Advanced Analytics, Content Analytics, Enterprise Search (or Information Discovery).
In the text analytics layer of our Big Data Project, business information and semantic contents are extracted from the textual big data mainly through statistical techniques and a domain dependent lexical analysis, rather than true semantic techniques.
An overall view of the above themes appears in Figure 1, where we show the component schema adopted for our company's model.

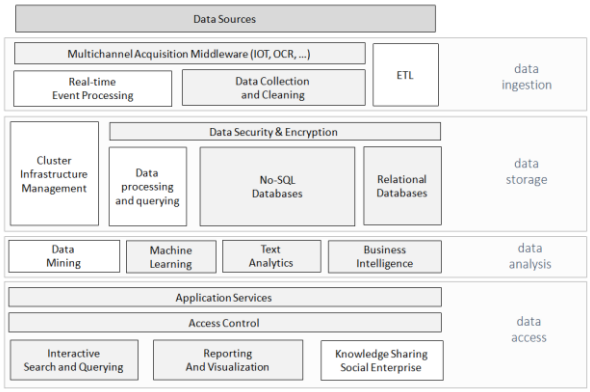

Figure 1. General overview of the technological components of the SIBDA System (components related to ECM have grey background)

Tools to manage large amounts of heterogeneous data in real time are increasingly important to address the new sources of data, such as devices and sensors of the IoT, social media web platforms, digital media and multimedia systems and business processes. The most significant techniques and technologies are discussed in [Bibri 2017]. "Variety" for Big Data means that data are in diverse formats, ranging from structured data to unstructured data (e.g., text documents). In addition, veracity and value are also critical characteristics of big data [Kune 2016]. Veracity refers to biases, noise, and abnormality in data. It is concerned with quality parameters such as uncertainty, unreliability, or inaccuracy of data.
In [Sandhu 2018], text analytics (or text mining) is presented as one of the major fields of Big Data, since most of the data are stored is in the form of text. Textual data can be unstructured due to different formats (e.g., Facebook posts, tweets, blogs, reports, files, and documents). These textual data contain many relevant pieces of information, which, if effectively analysed, can result in efficient decisions. Analyzing vast amount of textual data will lead to effective results in every business domain and will enable many industries to utilize this paradigm for interpreting their customers' needs.
In this paper, we consider also the growing interest in textual data mining. This interest calls for a coherent and intuitive method in BDA to discover hidden patterns and correlations in documents, and to monetize the data to improve the services, particularly for ECM.

3. IDENTIFYING INDICATORS

The automated ECM system developed by Microdata Group is framed in the BDA context both for the *volume* of textual data processed and for the *variety* and *variability* of handled documents.
First, we distinguish between process *efficiency* and *effectiveness*. The former is evaluated through a process quality assessment, carried out by a tracking software component, which measures the process performance. The latter is evaluated through an estimation of the data quality during the various processing stages, measured by different quality confidence indicators.
A problem emerged in the Project was the need to verify the *quality of data in the various steps* of the document management process. The issues regard the assessment of the effectiveness of the automation of each step, so that one can decide when to redirect the process to manual processing. In fact, currently, it is often necessary to perform a manual verification of the data recognized via automatic processing. It is necessary to monitor the quality of the overall work, to decide whether to send the document to the interactive residual processing, manually carried out by human operators.
Thanks to a measurement of the *estimated quality* of text, one can establish threshold values to decide the path that the document must follow (automated vs manual inspection and steps within each branch of the path, e.g. special tests). These threshold values may be changed and tuned based on various parameters, such as the document type, the customer, or the channel of document transmission (mail, web portal, traditional channel, and so on).
The relevance of evaluating the quality of data in the Project originates from the high heterogeneity of the textual sources of Big Data and from the poor quality of some document input channels (e.g., smartphone cameras). In fact, the data source of the textual data is a multichannel acquisition system

involving paper or files (containing the digitalized document image, or the document in various formats - e.g., pdf).
Therefore, we define three confidence indicators, evaluated at three successive points in the process, as shown in Figure 2. In general, confidence indicators used in our project are taken from the literature for their methods and computation (as summarized in a recent paper [Kiefer 2019]. These have been adapted to our case, computing them to give an estimation of the reliability each process step in the automated stages: i) OCR, ii) document classification, iii) data extraction.

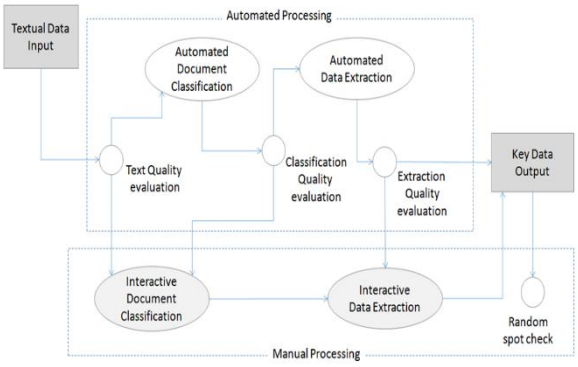

Figure 2. Computation of data quality indexes in the document processing workflow.

Each of the identified indicators assesses the quality of data from the previous stage and each has been approached with a different technique.

1) The first measure is the *Text Quality confidence* index, that is evaluated between the output of the texts from the OCR phase and the entry into the classification phase. It measures the quality of the textual data and therefore their reliability in view of the successive phases of automated processing. If the value is lower than a predetermined threshold, the document is conveyed to the interactive manual process performed by a human operator. In this case, also the subsequent key data extraction stage is carried out interactively by the human operator, both for practical reasons (he already visualized the document on his screen), and for the fact that, if the text quality is poor, the automated extraction is likely to lead to errors.
2) The second indicator, that we called *Classification Quality confidence index*, is evaluated at the exit of the classification stage, before entering the extraction stage. This is the simplest indicator to be computed, since the classification module based on ML already provides an uncertainty value of the classification result. This evaluation of uncertainty is combined with the value of the previous indicator, to decide whether to send the document to the automated data extraction or otherwise to send it to manual processing.
3) The third indicator is the *Extraction Quality confidence index* and is evaluated at the output of the key data automatic extraction stage. It is measured using a fuzzy lookup approach, based on support data provided by customers, to verify the quality of the extracted data. Through the fuzzy lookup technique, every data extracted from automatic processing is matched against a dictionary of valid words, including the personal data records of the specific customer. In this way, it is possible to identify the correct data, increasing the recognition rate of the index data and estimating a confidence index of the overall automatic processing.

We mainly focused on the *text quality confidence index*: the heterogeneity of the source channels implies the processing of low quality documents, due to geometric distortions or low resolution. A degraded text input greatly affects the subsequent process steps and for this reason we needed to implement a customized solution to check the quality of texts before sending the documents to automated classification and data extraction.

Therefore we developed a text quality evaluation component based on the combination of confidence estimation over successive text levels, starting from the XML output provided by the OCR (Optical Charcter Recognition) software. Through a series of nested cycles, we compute an estimated confidence value for each character, for each word and for each text line, converging in the overall document confidence index.

The measurement of the *character-level confidence* does not require specific techniques, as our OCR system directly provide a confidence value for each recognized character. The OCR also marks the portions of the document containing significant elements and identifies them by rectangular areas that are called bounding boxes. By setting a threshold value, we can estimate the total count of boxes believed to contain valid text and the count of boxes that probably contain invalid text. The corresponding ratio can provide a clue of the document deterioration degree.

Then we calculated the *word-level confidence*, by estimating the correctness of each token. A first estimation results from the average confidence at character level, integrated with the reliability

estimation of the bounding boxes and with the frequency of the words labeled as "suspicious" by the OCR software.

A more advanced check of a token plausibility can be based on lexical analysis techniques, based on the search in a vocabulary. First of all, we search in the vocabulary if the token is counted as a valid word. For very chaotic tokens, we select the word for which the string distance with the token is minimal is selected. A common solution is the Levenshtein distance, that measures the minimum number of variations needed to transform one string into the other.

The data generated by the OCR software explicitly provides the grouping of words in lines. This grouping is used to measure the confidence of the individual lines of text or *row-level confidence*, which can be used both as an indicator of the possible presence of deteriorated areas within the document, and as an intermediate step in the calculation of the overall document text quality.

The synthetic indicator that estimates the overall document-level text quality confidence index, results from a linear combination of the various factors described above, whose relevance is expressed by a series of weights. The value is then compared with a threshold, to determine whether the document can be sent to the automated classification and semantic information extraction process or it has to be manually processed by an operator. The algorithm can be tuned by modifying the parameters and threshold values, to identify the most effective configuration, according to the type of processing and the document category.

4. ARCHITECTURE AND TECHNOLOGY HINTS

The approach has been prototyped, providing for basis the experimentation of different specialized software tools, and applied in a defined context, based on Proofs of Concept (PoCs). The purpose was to evaluate the resulting quality, in terms of performances and error level, and to estimate the costs of adoption, both for software implementation and for the necessary infrastructure setup, before the subsequent step into production on a set of pilots.
The architectural design started from the need to extract information from documents even in the presence of a structure not described a priori formally, and to carry out a complete extraction of all the textual information contained in the document, without being limited to the information necessary for the required processing.
The implementation choices adhere to the following directions:

1) Choose components that can be integrated into an industrial process, within a server-based solution, avoiding the need to install software components on client stations and minimizing the operators' interventions.
2) Favour solutions capable of minimizing the specific implementation effort to be devoted to each processing or to each document type, exploiting tools with ML capabilities, which progressively improve the quality of the result as the number of cases processed increases. The overall software architecture of the ECM system, represented in the figure above, is based on a service-oriented architecture, supported by a service bus and by a SOA message broker, with the role of orchestrator, which invokes the different content classification and extraction modules that process the document.
The selection of this solution derives from two objectives:
i) to be highly scalable
ii) easy support changes in the technological tools, with limited impact on the process structure.
A process-tracking application performs both monitoring and supervision of the whole process and feeds an on-line dashboard that enables a fine-grained control of the automated jobs.
Considering the techniques used in the classification and indexing phases of documents, we distinguish between two portions of the architecture, namely:
- a portion dealing with analysis of documents with prevalence of text;
- a portion dealing with the analysis of documents with a prevalence of images.

In the classification phase of documents with prevalence of images, the classifier uses supervised ML algorithms based mainly on neural networks. The training set consists of images on which a pre-processing of the documents is performed (in particular a "cropping"), to maximize the effectiveness of the classification.
Regarding the classification of text documents, we apply a proprietary ML algorithm based on the Support Vector Machines (SVM) technique, which works on the weight assigned to a set of keywords. The system is enriched using text extracted from the training set documents, through which the algorithm learns the list of keywords and related weights useful for document classification.

The interactive human intervention is limited to the processing of the fraction of documents that have not been properly managed by the automated tools. Each manually elaborated document becomes a part of the training set of the ML engine, so that, at each subsequent step, the automatic systems has a higher probability of processing it.
The subsequent phase is the extraction of the characteristic meaningful information of each document, which starts from the document type returned by the classification phase and its geometric features. To maximize extraction yield, this stage focuses only on specific portions of the document.
For this data extraction task, different logics are applied, depending on whether the textual document is structured or unstructured.
For structured documents, a set of rules is used, employing:

- a recognition logic, based on regular expressions;
- the search for specific anchor elements in the documents.

For unstructured documents, a supervised ML algorithm is applied that operates as follows. In the preparation of the training set, tags are attached to the key information to be extracted. Then, the algorithm learns to recognize the textual patterns related to the occurrence of combinations of words and phrases, such as a statistic of the most frequent words that typically precede or follow the key information, without relying on a true semantic analysis. Finally, the algorithm extracts the relevant text portions from the document, for indexing purposes.
Upon completion of the classification and extraction phases, and after the possible document analysis interactive processing, the extracted information is stored in the database of the document management system.
For this purpose, the project uses a document-oriented database for Big Data, which allows for variable data structures according to the type of document. As for storage, the adoption of a NoSQL DB was selected, in particular a document-oriented DB, in order to store both the data extracted from the documents and the process data. Document-oriented DBs provide structured data types, typically in XML or JSON format (JavaScript Object Notation), which allow storing a multi-level text structure in a single point, thus facilitating the insertion and retrieval operations. For this purpose, the MongoDB DB server was adopted, which natively uses the JSON format, in an extended variant with binary encoding called BSON.
The complete textual data, i.e., the full text extracted from the document by OCR, are not currently saved, since this operation could possibly violate privacy issues. However, if a customer requires the availability of complete texts, for example for the design of new value-added services, the documents can be stored as text files on a dedicated file-server.

4.1 Experimental Results

After the PoC proved the technical feasibility of the automatic document classification and automatic key data extraction, the company set up a number of pilots for a first group of job process categories and experiments were carried out to test the automated process performance and to evaluate data quality.
The pilot tests concerned a few job categories, where it was easier to build a good document training set. It is necessary in effect to set up a dedicated training set for each processed document type, and although many document types are transversal to the different job processes categories, other are specific to each individual job category.
The resulting success rate is variable depending on the type of processed document and the pilot experiments are still continuing to gradually include new document types. The overall results here provided are therefore affected by the document types so far included in the training and they are also very sensitive to the input document acquisition channel.
For structured documents, such as contracts documentation or privacy consent forms, the classification levels considered as adequate for a practical use of the system have already been achieved.

5. CONCLUSIONS

The text analytics solution described in this paper shows how the calculation of data quality indexes along the big data analytics process has proved essential to obtain an effective result in document classification and semantic analysis practices.
We have presented an approach that combines three criteria for evaluating the quality of the content data. This combination allows one to avoid conveying good-quality texts to human operators. On the other side, it avoids to send to the automatic processing a document that contains degraded text, even when for example it is classified as “safe” by the automatic

classifiers and therefore with a high classification confidence value.

In the future, it will be possible to use i.e., two of the techniques introduced for the computation of quality indexes, namely, the text lexicality evaluation and the fuzzy lookup, to directly improve the quality of the texts, using them to perform automatic corrections.

As of performance evaluation, the use of metadata also makes it possible to detect the process tracking data in real time, and therefore to be able to access timely information on the work in progress. The collection of these data in an automated way, previously carried out manually by operators, makes it easier to analyse the load balancing and therefore to optimize the process.

Through dash boarding and reporting functions, precise estimates of processing times are possible. This increases future performance, thanks to smarter resource allocation. SLAs, that are so far negotiated on a prudential basis, could instead be evaluated more precisely. At the same time, the automated system made it possible to lower the level of jobs granularity from the document package to the level of a practice and soon to the single document, in order to have a more precise tracking of the process and a more flexible reallocation of resources.

Future work includes an automatic improvement of indicator computation, with feedback information coming from the classification and extraction steps of the same document, deriving from both automatic and manual processing. For example, by recovering and comparing processed texts with the key data extracted and with the classification results, it will be possible to contribute to the construction of contextual registries or to the identification of the critical areas of the documents.

Taking advantage of the capacity of the No-SQL documental database, Microdata has also started to save the full texts extracted from the documents (previously only the key data were saved) and the related confidence indexes, to later allow a number of off-line analyses on a significant volume of data. Among these future analyses, the tuning of the confidence index calculation parameters or the frequency estimation of words appearing in the different types of documents, to use both as a statistical model for verifying the text quality confidence and possibly for their automatic correction.

Building training sets that include two or three hundreds of documents having a high Text Quality Index score, and an accordingly low frequency of errors, we reach automatic classification rates between 90% and 95%.

For the identification documents, such as driving licenses and identity cards, it was necessary to build a larger set of training documents, reaching the thousand documents for each type, to achieve a successful classification rate around 90%.

As of the process phase of key data extraction, the resulting success rate are lower and they are more sensitive to the document type and to the quality of the original image of the digitized document. Rates are also affected by the wide variability of the document templates and contents, depending on the different customers.

Using the approach described in the sections above, the system achieved an average key data recognition rate of around 70% on more variable and less structured documents to 90% for more standardized ones.

In addition, the adoption of an automated system results in a dramatically cut of the document processing time. Hence, the number of documents processed in a unit of time increases by at least a factor of 15 for a single processing thread, to be multiplied for the level of parallel computation that could be deployed on the base of the available hardware infrastructures.

ACKNOWLEDGEMENTS

We are thankful to the three companies in the Big Data Project: MailUp, one of the leading Italian Email Service Providers, Microdata Group, specialized in outsourcing services for management of document processes, and Linea Com, operating in the IT and telecommunications services.

REFERENCES

[Batini 2016] Batini C, Scannapieco M. Data and information quality. Cham, Switzerland: Springer International Publishing. 2016.

[Bibri 2017] Bibri, S.E. and Krogstie, J., 2017. ICT of the new wave of computing for sustainable urban forms: their Big Data and context-aware augmented typologies and design concepts. Sustainable cities and society, 32, pp.449-474.

[Chawda 2016] Chawda, R.K. and Thakur, G., 2016, March. Big data and advanced analytics tools. In 2016 Symposium on Colossal Data Analysis and Networking (CDAN) (pp. 1-8). IEEE.

[Fugini 2018] Fugini, M. and Finocchi, J., 2018, June. Innovative Big Data Analytics: A System for Document Management. In 2018 IEEE 27th International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE) (pp. 267-274). IEEE.

[Fugini 2019] Fugini M, Finocchi J, Leccardi F, Locatelli P, Lupi A. A Text Analytics Architecture for Smart Companies. In 2019 IEEE 28th International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE) 2019 Jun 12 (pp. 271-276). IEEE.

[Kiefer 2019] Kiefer, C., 2019. Quality indicators for text data. BTW 2019–Workshopband.

[Kune 2016] Kune, R., Konugurthi, P.K., Agarwal, A., Chillarige, R.R. and Buyya, R., 2016. The anatomy of big data computing. Software: Practice and Experience, 46(1), pp.79-105.

[Marjani 2017] Marjani, M., Nasaruddin, F., Gani, A., Karim, A., Hashem, I.A.T., Siddiqa, A. and Yaqoob, I., 2017. Big IoT data analytics: architecture, opportunities, and open research challenges. IEEE Access, 5, pp.5247-5261.

[Moniruzzaman 2013] Moniruzzaman, A. B. M., and Syed Akhter Hossain. "Nosql DB: New era of DBs for big data analytics-classification, characteristics and comparison." arXiv preprint arXiv:1307.0191 (2013).

[Nicola 2014] Nicola, S.; Ferreira, E.P.; and Ferreira, J.J.P. A quantitative model for decomposing and assessing the value for the customer. Journal of Innovation Management, 2, 1 (2014), 104–138.

[Qiao 2017] Qiao, Z., Zhang, X., Zhou, M., Wang, G.A. and Fan, W., 2017, January. A domain oriented LDA model for mining product defects from online customer reviews. In Proceedings of the 50th Hawaii International Conference on System Sciences.

[Sandhu 2018] Sandhu, Rajinder, Jaspreet Kaur, and Vivek Thapar. "An effective framework for finding similar cases of dengue from audio and text data using domain thesaurus and case base reasoning." Enterprise Information Systems 12.2 (2018): 155-172.

[Wang 2018] Wang, Y., Kung, L. and Byrd, T.A., 2018. Big data analytics: Understanding its capabilities and potential benefits for healthcare organizations. Technological Forecasting and Social Change, 126, pp.3-13.