Quality Evaluation for Documental Big Data

Mariagrazia Fugini and Jacopo Finocchi
Dipartimento di Elettronica Informazione e Bioingegneria, Politecnico di Milano, Piazza L. Da Vinci 32, Milano, Italy

Keywords: Text Analytics, Big Data Analytics, Enterprise Content Management, Document Management, Machine Learning for Document Processing.

Abstract: This paper presents the analysis of quality regarding a textual Big Data Analytics approach developed within a Project dealing with a platform for Big Data shared among three companies. In particular, the paper focuses on *documental Big Data*. In the context of the Project, the work presented here deals with extraction of knowledge from document and process data in a Big Data environment, and focuses on the *quality* of processed data. Performance indexes, like correctness, precision, and efficiency parameters are used to evaluate the quality of the extraction and classification process. The novelty of the approach is that no document types are predefined but rather, after manual processing of new types, datasets are continuously set up as training sets to be processed by a Machine Learning step that learns the new documents types. The paper presents the document management architecture and discusses the main 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 (Singh, 2019; Wang, 2018). In particular, we are witnessing a growing relevance of *unstructured information* for Enterprise Architectures, compared to structured data stored in traditional database systems. The spread of Big Data has changed the type of data collected and processed by information systems and is made more and more suitable to unstructured, large data, such as images, streaming data, enterprise documents, and so on.

This paper describes a solution for *document processing automation* based on *Enterprise Content Management (ECM)* in the context of Big Data. The focus of the paper is on the definition of *quality indexes* enabling to evaluate the performance, precision, effectiveness, and other quality features of BDA for document management. It presents an experimental BDA approach in a real setting, given by the studies performed in the *SIBDA* – Sistema Innovativo di Big Data Analytics - Project [1] concerning an innovative solution to Big Data collection, analysis and management in a cooperative shared enterprise environment. This Project develops a *solution shared among three companies*, structured in a temporary agreement for competitiveness.

SIBDA integrates the activities and assets of the three involved companies, namely:

- *MailUp*, leader of the project, one of the main Italian operators in the field of Email Service Providers; it developed the BD storage infrastructure and provided the general-purpose BDA tools.
- *Microdata Service*, a company specialized in outsourcing services for management of document processes; it developed automatic document analysis tools to extract information from heterogeneous and unstructured data sources.
- *Linea Com*, operating in IT and telecommunications services in the area of fiber-optic and wireless networks for various municipalities of Regione Lombardia; it

[1] *Big Data Analytics Project - SIBDA*, funded by Regione Lombardia (201-2018) Progetto Competitività Cremona, "Smart Cities", see *IEEE WETICE Best paper* IEEE WETICE 2018 - W2T Conf. (Paris, June 27-29, 2018) Available at: https://scholar.google.com/scholar?as_ylo=2016&q=fugini+wetice+2018&hl=en&as_sdt=1,5

Fugini, M. and Finocchi, J.
Quality Evaluation for Documental Big Data.
DOI: 10.5220/0009394301320139
In *Proceedings of the 22nd International Conference on Enterprise Information Systems (ICEIS 2020) - Volume 1*, pages 132-139
ISBN: 978-989-758-423-7

created a platform for the acquisition of sensor data.

In the area of BDA, *SIBDA* tackled Enterprise ECM (Shivakumar, 2016; Vlad, 2019). ECM for documents is the subject of this paper, that we studied in cooperation with Microdata Service.

As an evolution of traditional document management applications, ECM moves towards advanced content management applications, characterized by the growing significance of semantics in the enterprise environment and by the diffusion of Big Data tools. In such scope, a relevant trend in the analysis of unstructured data is the application of Natural Language Processing (NLP) techniques - frequently based on *Machine Learning (ML)* - to draw semantic value from textual data and use the semantics to enhance search functionalities or to develop new value-added services. The studied ECM is inspired to a context-aware computing model: document semantics is extracted based on a knowledge of the process where they are immersed (Fugini, 2019).

In this paper, the focus is on incorporating *data quality parameters* and *process quality features*, such as, efficiency, effectiveness, accuracy, etc., in *management processes involving Big Data Documents*. The introduction of these parameters can provide the capability of evaluating service performance and textual data quality (Batini, 2016; Kiefer, 2019).

The paper is organized as follows. Section 2 reviews the *SIBDA* Project and the approach to Quality Indexes. Section 3 focuses on indicators of *process* quality and *data* quality. Section 4 describes the system architecture and technologies, commenting basic results. Section 5 concludes the paper.

2 PROJECT AIMS AND QUALITY INDEXES

The aim of the *SIBDA* Project is 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. The overall research aim is 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. In (Fugini, 2018; Fugini, 2019) we give details about the overall Project. The portion of Document Big Data illustrated in this paper is concerned with private data owned by Microdata regarding customers.

In particular, the goals of *SIBDA dealing with documents* are twofold: industrial and scientific.

Industrial aims can be summarized as follows:

1. to reduce the human work involved in document management with potentially no loss of information nor of correctness during the data extraction process;
2. to increase the volume of processed documents per time unit, while respecting aim 1 above;
3. to enable going beyond key data extraction, namely to enable storing the whole document text for analysis in the light of making proposal of new services to the customers.

Scientific aims can be summarized as follows:

1. to classify texts with a good precision rate, with *no need* to define a *template* for new document types, instead using ML to learn the structure of new documents;
2. to make some indicators available to understand automatically whether a document should be forwarder to *manual* or to *automatic* classification. This allows the system to classify manually or automatically the input documents, depending on their quality; in fact, documents can enter the system via fax, via smart phone, as full digital texts or as images, and so on, with different quality.

2.1 Quality Indexes

We identify indexes of quality for assessment and for performance monitoring of *textual BDA*. We illustrate the quality indexes, the system architecture, and the basic results achieved by Politecnico di Milano and Microdata Group[2].

When building a BDA system, core data assets present many challenges, such as data quality, data integration, data analysis on metadata, and so on. In particular, data quality problems add complexity to the use of Big Data, with several general data quality challenges.

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

First, we observe that 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.

Secondly, we point out that, as data grow exponentially, it becomes increasingly difficult for companies to ensure that their sources of data and information therein 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 (Qiao, 2017; Hayek, 2020). In our Project, *data validity,* more than an issue of *veracity,* is a parameter depending on the data quality at the input point.

Given these premises, data cleaning, filtering, and selection techniques able to detect and remove noise and abnormality from data automatically become essential.

In any knowledge extraction process, the value of extracted knowledge is related to the quality of the used data. Big Data problems, generated by massive growth in the scale of data observed in recent years, also follow the same issue. A common problem affecting data, and Big Data, quality is the presence of noise, particularly in classification problems, where label noise refers to the incorrect labeling of training instances, and is known to be a very disruptive feature of data (Garcia-Gil, 2019). In our approach, data are labelled initially by human operators, and gradually passed on as a training set to the supervised ML algorithms. This procedure avoids having many partially or even unlabelled data.

Our research focuses on the development of uniform data quality standards and metrics for BDA that address some of the various data quality dimensions (e.g., accuracy, accessibility, credibility, consistency, completeness, integrity, auditability, interpretability, and timeliness). In particular, our research focuses on the set up of a *panel of indicators* for the analysis of *performance* and *quality aspects* of the *BDA process*.

To this aim, the most significant techniques and the technologies considered in *SIBDA* belong to 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 regards 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, including the whole documents; ii) how to archive unstructured and variable data in a way that makes them understandable automatically via a ML approach. The solution lies in adopting NoSQL, or document-oriented databases.

Another issue related to storage regards 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 (Meier, 2019). 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, defined as *Advanced Analytics* (Chawda, 2016; Elragal, 2019). The generated knowledge has to be available to users and shared effectively among 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) (Hariri, 2019).

An overall view of the above themes is in Figure 1, where we show the component schema adopted for our company's model.

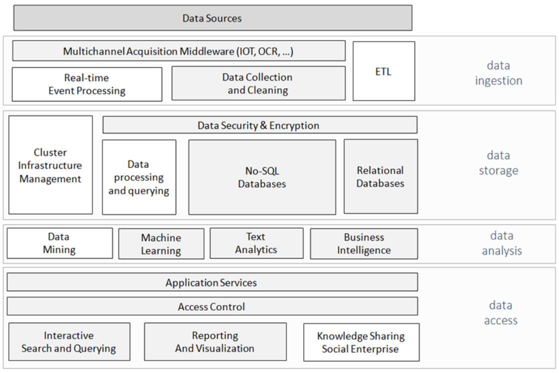
Data Sources
Multichannel Acquisition Middleware (IOT, OCR, ...)
Real-time Event Processing
Data Collection and Cleaning
ETL
data ingestion
Data Security & Encryption
Cluster Infrastructure Management
Data processing and querying
No-SQL Databases
Relational Databases
data storage
Data Mining
Machine Learning
Text Analytics
Business Intelligence
data analysis
Application Services
Access Control
Interactive Search and Querying
Reporting And Visualization
Knowledge Sharing Social Enterprise
data access

Figure 1: General overview of the technological components of the *SIBDA* System (components related to ECM are in grey background).

3 IDENTIFYING INDICATORS

The automated ECM system developed by Microdata Group and DEIB-Politecnico di Milano in *SIBDA* is framed in the BDA context both for the *volume* of textual data processed and for the *variety* and *variability* of handled documents.

To define quality indicators, we first have to 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.

During the project, we ascertained the need to verify the *quality of processed data* not only at the end of the process, to assess the adequacy of the returned results, but also *in the various steps* of the content management process. A data quality assessment had to be embedded in the process in order to be able to route the data to the most appropriate sub-process. In fact, currently, it is often necessary to decide whether to send the document to the so-called "residual interactive 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. specific tests). These threshold values may be set and tuned according to various parameters, such as the processing job type, the document type, the customer, or the channel of document transmission (paper, e-mail, web portal, mobile app, and so on).

The relevance of evaluating the quality of textual data in the Project originates from the high heterogeneity of our Big Data document sources 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 various document and file formats.

Therefore, we define *three confidence indicators*, evaluated at three successive points in the process, as shown in Figure 2.

As a basis, confidence indicators used in our Project are taken from the literature, since their methods and computation are suitable, as summarized in a recent paper (Kiefer, 2019). These have been adapted to our case, computing them to give an estimation of the reliability at each process step in the automated stages: i) OCR, ii) document classification, iii) data extraction.

Each of the identified indicators assesses the quality of data from the previous stage, approached with a different technique, as illustrated in the following.

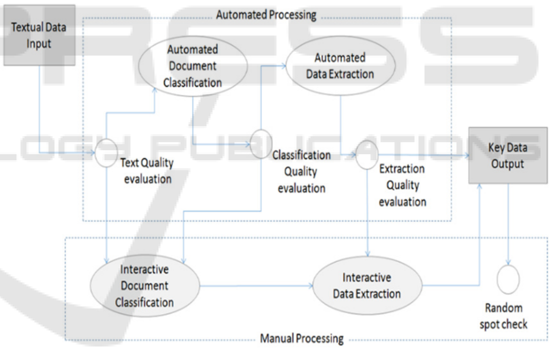
Textual Data Input
Automated Processing
Automated Document Classification
Automated Data Extraction
Text Quality evaluation
Classification Quality evaluation
Extraction Quality evaluation
Key Data Output
Interactive Document Classification
Interactive Data Extraction
Random spot check
Manual Processing

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

1. The first measure is the *Text Quality confidence index*, which is evaluated by comparing the output of the texts from the OCR phase with 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/she already visualized the

document on the 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 call *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 (Joachims, 2002).

3. The third indicator is the *Extraction Quality confidence index*, which 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 (Bast, 2013).

In the first phase of the Project, 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 Character 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 computed 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 labelled as "suspicious" by the OCR software.

In the subsequent phases of the Project, a more advanced check of a token plausibility has been developed, based on *lexical analysis* techniques, using a search within a vocabulary. First, we search in the vocabulary if the token is a valid word. For very ‘chaotic’ tokens, we select the word for which the string distance with the token is minimal, implementing a common *weighted edit distance* measurement (a generalization of the Levenshtein distance).

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.

To compute the synthetic indicator that estimates the overall document-level *text quality confidence index*, we perform a linear combination of the various level confidence 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 is tuned by modifying the parameter weights and the 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 (PoC). The purpose was to evaluate the resulting process 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.

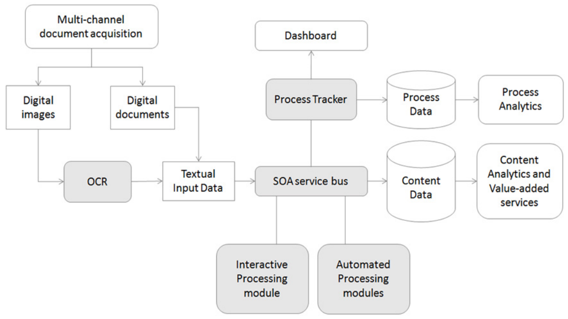
Multi-channel document acquisition
Digital images
Digital documents
OCR
Textual Input Data
SOA service bus
Dashboard
Process Tracker
Process Data
Process Analytics
Content Data
Content Analytics and Value-added services
Interactive Processing module
Automated Processing modules

Figure 3: ECM system architecture.

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 software components that can be integrated into an industrial process, ensuring scalability and high availability requirements compliance, while avoiding the need to install 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 job or to each document type, given the frequent changes in the content and structure of the documents and the growing number of document types to be processed.

The first requirement is met by a sever-based solution, deployed on a hyper-converged infrastructure (based on a commercial HCI platform). In particular, the overall software architecture of the ECM system, represented in Figure 3, 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.

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.

The second requirement is fulfilled by exploiting tools with ML capabilities, which progressively improve the quality of the result as the number of processed documents increases, and avoiding the solutions based on building a specific template for each document type to process.

Regarding the classification step of text documents, we therefore 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 process step is the extraction of the characteristic content meaningful information of each document, which starts from the document type returned by the classification step 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.

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 content extracted from the document aimed for example to the design of new value-added services, currently is not saved, since the possible violation of privacy issues is still under discussion.

As of process performance evaluation, the use of metadata attached to each document 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, which 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.

4.1 Discussion

When 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. 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, we provided training sets that include two or three hundreds of documents having a high *Text Quality Index* score, and an accordingly low frequency of errors, reaching 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. In fact, 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 by the level of parallel computation that could be deployed on the base of the available hardware infrastructures.

5 CONCLUDING REMARKS

The text analytics solution for documents described in this paper shows how the calculation of data quality indexes along a Big Data Analytics process can give effective results 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, we plan to use a combination 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.

REFERENCES

Bast, H., & Celikik, M. (2013). Efficient fuzzy search in large text collections. *ACM Transactions on Information Systems (TOIS)*, *31*(2), 1-59.

Batini, C., & Scannapieco, M. (2016). Data and information quality. *Cham, Switzerland: Springer International Publishing. Google Scholar*, 43..

Chawda, R. K., & Thakur, G. (2016, March). Big data and advanced analytics tools. In *2016 symposium on colossal data analysis and networking (CDAN)* (pp. 1-8). IEEE.

Elragal, A., & Hassanien, H. E. D. (2019). Augmenting advanced analytics into enterprise systems: A focus on post-implementation activities. *Systems*, *7*(2), 31.

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

García-Gil, D., Luengo, J., García, S., & Herrera, F. (2019). Enabling smart data: noise filtering in big data classification. *Information Sciences*, *479*, 135-152.

Hariri, R. H., Fredericks, E. M., & Bowers, K. M. (2019). Uncertainty in big data analytics: survey, opportunities, and challenges. *Journal of Big Data*, *6*(1), 44.

Hajek, P., Barushka, A., & Munk, M. (2020). Fake consumer review detection using deep neural networks integrating word embeddings and emotion mining. *Neural Computing and Applications*, 1-16.

Joachims, T. (2002). *Learning to classify text using support vector machines* (Vol. 668). Springer Science & Business Media.

Kiefer, C. (2019). Quality indicators for text data. *BTW 2019–Workshopband*.

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

Meier, A., & Kaufmann, M. (2019). Nosql databases. In *SQL & NoSQL Databases* (pp. 201-218). Springer Vieweg, Wiesbaden.

Qiao, Z., Zhang, X., Zhou, M., Wang, G. A., & Fan, W. (2017). A domain oriented LDA model for mining product defects from online customer reviews.

Shivakumar, S. K. (2016). *Enterprise content and search management for building digital platforms*. John Wiley & Sons.

Singh, S. K., & El-Kassar, A. N. (2019). Role of big data analytics in developing sustainable capabilities. *Journal of cleaner production*, *213*, 1264-1273.

Vlad, M.P. and Mocean, L., 2019. About document management systems. Quaestus, (15), pp.217-225.

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