Big Data Quality Assessment Model for Unstructured Data

Ikbal Taleb
CIISE
Concordia University
Montreal, QC, Canada
i_taleb@live.concordia.ca

Mohamed Adel Serhani
College of Information Technology
UAE University
Al Ain, UAE
serhanim@uaeu.ac.ae

Rachida Dssouli
CIISE
Concordia University
Montreal, QC, Canada
rachida.dssouli@concordia.ca

***Abstract*— Big Data has gained an enormous momentum the past few years because of the tremendous volume of generated and processed Data from diverse application domains. Nowadays, it is estimated that 80% of all the generated data is unstructured. Evaluating the quality of Big data has been identified to be essential to guarantee data quality dimensions including for example completeness, and accuracy. Current initiatives for unstructured data quality evaluation are still under investigations. In this paper, we propose a quality evaluation model to handle quality of Unstructured Big Data (UBD). The later captures and discover first key properties of unstructured big data and its characteristics, provides some comprehensive mechanisms to sample, profile the UBD dataset and extract features and characteristics from heterogeneous data types in different formats. A Data Quality repository manage relationships between Data quality dimensions, quality Metrics, features extraction methods, mining methodologies, data types and data domains. An analysis of the samples provides a data profile of UBD. This profile is extended to a quality profile that contains the quality mapping with selected features for quality assessment. We developed an UBD quality assessment model that handles all the processes from the UBD profiling exploration to the Quality report. The model provides an initial blueprint for quality estimation of unstructured Big data. It also, states a set of quality characteristics and indicators that can be used to outline an initial data quality schema of UBD.**

Keywords—Big Data, Data Quality, Unstructured Data, Quality of Unstructured Big Data.

I. INTRODUCTION

Big data is commonly defined as the way we gather, store, manipulate, analyze and get insight from a fast-increasing heterogeneous data. Most of the new generated data is unstructured due to the increase of mobile and human's unlimited generated data from social medias that combine text, pictures, audio, video, in an unstructured way. Unstructured data is a fast-increasing phenomenon than all other types of data, industry analysts say. It will increase by as much as 800 percent during the next five years according to a survey conducted by [1]. This urge the need to automatically characterize and categorize such data. These classifications are strongly coupled with the semantic meaning of what the data represents. In many cases, the data comes in a format and a quality state in which it is impossible to process immediately as it is, and if so, the results cannot guarantee a valuable analysis and insights.

Big Data Quality assessment is an important phase integrated within data pre-processing. It is a phase where the data is prepared following the user or application requirements. When the data is well defined with a schema, or in a tabular format, its quality evaluation becomes easier as the data description will help mapping the attributes to quality dimensions and set the quality requirements as baseline to assess the quality metrics. In the other case, when there is no structure to follow, an intermediary phase needs to be defined to parse, analyses, mine, detect a schema or a path the unstructured data went through. To achieve this objective, a set of techniques such as classification, clustering, searching or mining is used to draw a set of artifacts that act as a filter, a translator of UBD to a more efficient readable format ready for quality evaluation. The amount of resulting data is generally equal or far less than the input. A reduction of managed data is to be defined to assess the efficiency and impact of these techniques on the intermediate assessment results.

Assessing the quality of unstructured data is a tedious task. In the following, we enumerate some data characteristics and data quality aspects that adds more difficulty to the assessment process: (1) data size, (2) heterogeneity (3) multiple data type and formats (4) multi-sources, multi-files (5) what DQD to choose from? (6) define clearly the quality of unstructured data, for example what are the quality dimensions for a UBD set contain: Text files, Images, Videos, Audio files, Web pages, PDF files, twitter data, Facebook data, etc. In a presence of such data diversity, we need to define the followings:

a) UBD Quality Project
b) Set of requirements with default DQD's to start with.
c) Sampling strategy for UBD that must consider the type and format of the data. There is no attributes or observations that can be used as the basis for sampling
d) How to extract features, variable, attributes that characterize UBD, and evaluate its quality based on these finding.
e) Select the best techniques, methods, strategies that extract a useful information from UBD or convert it to a schema-based data.

In the age of Big Data, many trending data analytics directions are now focusing on the analysis of customer behavior, feedback, comments about their products or services. They mine the social media data streams, from Twitter, Facebook, YouTube, Instagram, websites, forums, text messaging to get some valuable insights. A thousand of terabyte of data is available to be analyzed using techniques such as

978-1-5386-6673-9/18/$31.00 ©2018 IEEE

Authorized licensed use limited to: R V College of Engineering. Downloaded on June 29,2024 at 08:57:41 UTC from IEEE Xplore. Restrictions apply.

sentiments analysis, and deep learning. The necessity to assess the quality of this data before engaging in large processing that costs time and money is a must. Approximating the quality of such data sets is the first step towards successful big data project. Finally, structured data is always easy to be handled as Big Data by data analytics applications rather than unstructured one.

The rest of paper is organized as follow: next section introduces Big Data and data quality fundamentals, definition, characteristics, and lifecycle. Section 3 surveys the most important research works on Unstructured Big Data quality evaluation and management. Section 4 introduces our Unstructured Big Data quality assessment Model. Section 5 analyses and discuss our model experimentation's. Finally, the last section concludes the paper and points to some ongoing and challenging directions.

II. BIG DATA FUNDAMENTALS

In this section, we introduce some Big Data foundations and all the elements that cooperate to contribute to such ecosystem.

A. Big Data overview

We always dealt with Big Data, the moment we started gathering data and storing it in different ways. Big Data is being considered in every domain, in academia, in industry, in businesses, in social media, and in research. It has a lifecycle and characteristics to be defined and followed.

1) Big Data Lifecycle

Figure 1 describe the most important stages that the data goes through till the purpose that it was gathered and used for. From the data inception, collection, transport through inter-networks, saved into distributed storage around the world that offers the best quality price with a reliable network. Then pre-processed to filter only the best quality data and forwarded to processing and analytics for insight extraction.

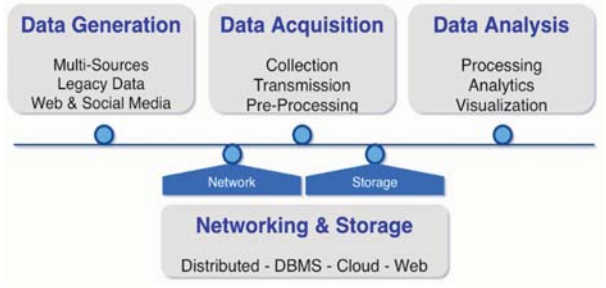

Fig. 1. Big Data Lifecycle

2) Big Data Characteristics

In the annual McKinney Global Institute report [2], three data dimensions characterizing Big Data were introduced. The Volume, Velocity and Variety, also called the 3 V's. Lately the number of dimensions increased from 3, 4, 7 and even to 10 V's [2]–[7]. As illustrated in Table 1, we compiled the most important V's that describes Big Data. As the name suggest Big Data is more than simply a matter of size; it is a prospect to unearth insights to make beneficial decisions. Thus, Visualization, Variability, Volatility, Virality, Vulnerability, Viscosity, Validity are extended characteristics.

B. Unstructured Big Data

To make decision we need relevant information that is extracted from data using processing and analysis. In this rich context, data exist in numerous formats, with different types and from several sources and knowledge domains. Unstructured data is growing faster than structured data. It is explained by the number of Facebook posts, tweets, photos and emails created in every second.

Table 1. The 5 V's characteristics of Big Data

#	Big Data V's	Description	Attributes & Metrics
1	**Volume**	Scale and Size of Data in Storage	Terabyte to Exabyte
2	**Velocity**	Data generation frequency: the speed in which this data is generated, produced, created, refreshed, and streamed. Viscosity: how difficult is the data to work with?	Milliseconds to seconds. Batch, Near Time, Real Time, Streams.
3	**Variety**	Multiple different forms of the data	Binary, raw, text, multimedia. Structured, **unstructured**, and semi-structured.
4	**Veracity**	Uncertainty of the Data that leads to confidence or trust in the data. How can we trust the data? What is its provenance? Is it reliable? Is it accurate? Is the data verifiable and truthful? What is Rigor in Data analysis?	Inconsistency, incompleteness, ambiguity, latency, trustfulness, and traceability (provenance).
5	**Value**	Deriving business value and insights from the data.	Big Data strategy, Big Data project, targets goals and suitable analytics process.

1) Unstructured Data

By default, the name unstructured data imply mess, noise and a chaos in data organization. In contrast, it refers to a data that doesn't have a schema, no Metadata, and no rules or constraints to follows when it has been created; likewise, the structural database model. Even it has some basic low-level internal structure but no pre-defined data models or schema. Unstructured data has two meaning: (1) no structure at all or (2) an unknown structure. It may be textual or non-textual, and human or machine-generated. It may also be stored within a non-relational database like No-SQL. In Table 2, we illustrate some unstructured data domains and the data types it generates and manages.

Table 2. Unstructured Data Domains

Data Domains	Data types
Healthcare	Doctors notes, X-rays, IRM, scanner images
Finance	Stock market data, bank transactions
Scientific Research	DNA data, satellite data
Customer Relationship Management (CRM)	Customer feedback, forums comments
Social Media	Facebook posts, twitter,
Media contents	Videos, images, audio, speech, music
Web Contents	Web pages, blogs, news
IoT	Sensor data, RFID data
Log files	Network logs, web pages click, Facebook logs
Documents	text, web, pdf, office docs, scanned docs

2) Unstructured Big Data Characteristics

In addition, to the noticeable difference which is the columnar data model, the major difference is the effortlessness of analyzing structured data versus unstructured data. The existing pre-processing and analytics tools are very mature for structured

Authorized licensed use limited to: R V College of Engineering. Downloaded on June 29,2024 at 08:57:41 UTC from IEEE Xplore. Restrictions apply.

data, but still in embryo state for unstructured one. The Variety characteristic of Big Data defines different formats of data (e.g. document, emails) that are not always stored in structured relational database systems. Its follow two classes of UBD:

- *Human generated*: Text files, Emails, social media, websites, mobile data messages, messaging chats, business applications, Media files (audio, video, image)
- *Machine generated*: scientific data, satellite imagery, digital surveillance, sensor data, network logs, IoT devices.

Many different sources of data in several domains that feeds the unstructured contents which prevail the name of data domains illustrated in Table 3. Therefore, it is important to note that unstructured big data is also characterized by velocity and volume.

3) Unstructured Data Management

Since big data can include both structured and unstructured data, the exploration of unstructured data can be handled using existing Big Data Ecosystems. Such systems and tools include Hadoop, business intelligence software (analytics, data mining, reporting), Data Integration Tools, Document Management Systems, Search and indexing Tools, Unstructured Information Management Architecture (UIMA) [8] and IBM/Apache component software architecture for analysis of unstructured data. In the following, we highlight some unstructured data types and some methodologies used to manage and analyze them.

a) Textual (text, Pdf, scanned docs, email body)

Unstructured textual data is transformed and explored using a combination of techniques as in Text Mining [9]–[12] such as data mining, machine learning, Natural Language Processing (NLP), information retrieval and knowledge management. Moreover, Search engines tools are used for indexing, cataloguing, categorizing to make information and text search easy characterizing the Unstructured Data. Also, other Techniques are used varying from text analytics, OCR to patterns, terms, topics detection and discovery for the sake of structuring the textual data.

b) Social Media (Twitter, Facebook), CRM

For twitter data, sentiment analysis [13]–[15], opinion mining [16]–[18], are well-known techniques applied to extract trends in multitude of areas like elections, events and much more. In CRM systems, a semantic analysis on multisource unstructured data a semantic analysis is conducted to annotate, extract, and rate customer feedbacks.

c) Media (Video, Audio, Image)

Digital photos, Videos, and Audio files are stored in a structured format such as JPG/ PNG, Mov/MP4, and WAV/MP# respectively. However, all these data don't express any information about what is in the data. It needs to be treated to comprehend its meaning. Automatic Media data tagging, labelling, indexing after analyzing and processing will help to search within the media files efficiently. Processing this kind of unstructured data needs some advanced algorithms for image, audio, speech, and video processing to gather patterns or any information that can be indexed.

C. Data Quality (DQ)

According to [19], data quality is not easy to describe, its meanings are data domain dependent and context-aware. Overall, data quality is continuously related to the quality of its data source [20].

Data Quality is differently perceived in both academia and industry. In [21], data quality from ISO 25012 Standard is defined as "the capability of data to satisfy stated and implied needs when used under specified conditions". However, in [22], it is summarized as "fitness for use" or "meeting user needs".

1) Data Quality Dimensions (DQD's)

According to [22]–[24], to measure and manage data quality the concepts of Data Quality Dimension (DQD) is presented. There are many quality dimensions that are classified under categories that define them. In Figure 4, we summaries some essential DQD categories based on [12]–[14]: the contextual dimensions that are associated to the information and intrinsic dimensions that refer to objective and native data attributes. Examples of intrinsic data quality dimensions include Accuracy, Timeliness, Consistency, and Completeness. Each Data Quality Dimension is associated with a specific metrics. A metric is a method, or formula established to measure a score or ratio from the data by quantifying its DQDs. A metric provides how to evaluate a DQD from simple formulas to more complex multivariate expressions.

2) Data Quality Assessment

With a set of Metrics, it is possible now to evaluate quantitatively the quality when following a data driven strategy on existing data. For structured data, its quality assessment is apparent as data is available and attributes with their corresponding values are accessible. However, for unstructured data, needs a different approach when we don't know how it is organized, and what are we are going to assess. The introduction of a module that extract, discover, or define attributes and features with specific DQD mapping is mandatory to proceed with the quality exploration.

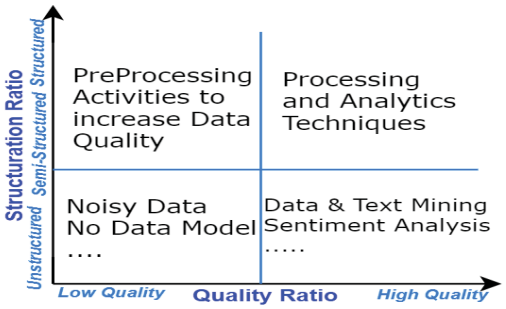

Fig. 2. Data Quality and Data Structure

3) Unstructured Data Quality

We initially should ask the following question: "*What and how should we measure, evaluate or assess in this big diversity of heterogeneous unstructured data*?". First, we need to notice that not all DQD's apply here, since most of them are used for structured data. Even if they are applicable to all, some intermediate dimensions might be defined for each type of unstructured data. For example, readability of text data is assessed by its reading easiness [25]. Then, transforming it to a structured data that can be measured and queried. For example,

Authorized licensed use limited to: R V College of Engineering. Downloaded on June 29,2024 at 08:57:41 UTC from IEEE Xplore. Restrictions apply.

in Text mining, a combination of techniques such as data mining, machine learning, Natural Language Processing (NLP), information retrieval and knowledge management are used to map the data to a schema. The quality assessment of the extracted relevant features that explains some quality indicators will results in a set of quality scores that can be easily mapped to traditional DQD's [24], [26], [27]. So, the quality of the unstructured data will depend on the structure that the data will fit into. Figure 2 provides an illustration about the relationship between data quality and data structure. The more the data is structured the more its quality increases and its evaluation becomes easier.

In [28], the authors extracted quality indicators from patient charts using a quantitative approach. They used canary, an NLP software to discover and mine knowledge hidden in unstructured clinical data. While authors in [29], [28], [30] used quality indicators but specifically for unstructured text data mining processes such as Interpretability, Relevancy, and Accuracy. The exploitation of these indicators must be converted into a number in the range of [0,1] as expressed by quality metrics to measure a structured DQD (e.g. completeness).

D. Quality of unstructured Big Data

A data driven strategy is followed to handle the quality of unstructured data. A set of steps and settings or pre-requirements must be defined before proceeding with the Quality Assessment of the Unstructured Big Data. If the Structured data model makes the quality assessment process relaxed with a set of known columns (attributes) organized in rows (observations). It is not the case for unstructured data, which includes many intermediate processes or modules to either 1) convert the data to a structured one and assess its quality or 2) use new techniques to extract meaningful features that represents the data and apply quality evaluation methodology. We illustrate in Figure 3, the steps we need to proceed to accomplish the quality assessment of Unstructured Big Data.

1) ***What is the Type, Format, or Data Domain*?** discovering these information's or extracting it from metadata or any description that came with data is priceless since it is essential to start with this process to explore the data contents.
2) ***What DQDs to use to map the Quality*?** depending of the data type, the DQD's are selected and then mapped to the unstructured data indicators. Even, the effectiveness of DQD selection is related to the discovery of attributes and features that are mapped to DQDs.
3) ***What Quality Metrics to consider*?** with unknown data features or new features, a creation, update, rewrite, or fork of exiting metrics to handle new discovered or extracted data features is mandatory. For example, the contrast ratio of an image is obtained by a metric that reads the digital picture and compute contrast intensity % using the related formula [25]. In [31], [32] more metric and techniques are enumerated to assess the quality of multimedia data.
4) ***How to identify attributes or features to evaluate*?** a list of attributes can be discovered easily for certain format and data types. Other feature extraction algorithms based on data domain, format and types are also needed.
5) ***What Sampling strategies to use*?** with large volumes of unstructured data, the quality estimation of a representative population is mandatory as we don't want the time and cost of preprocessing especially for unknown unstructured data to explode big data project budget.

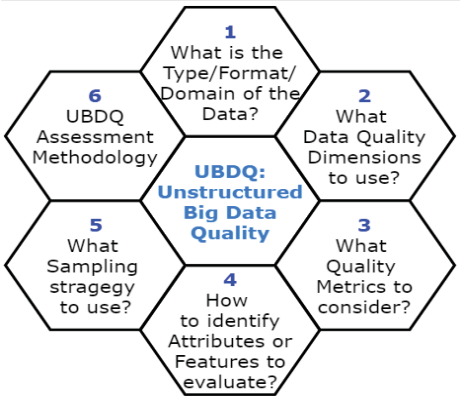

Fig. 3. Unstructured Big Data Quality Proceeding

6) ***What Quality Assessment Methodology to consider?*** the quality assessment for each tuple (DQD, feature, Samples) is representative of the whole Data; the choice of data sample size and iterations depend on the sampling strategy used.

III. LITERATURE REVIEW

We survey in this section, the very few available works on unstructured data quality assessment, UBD management and exploration and we highlight at the end the remaining challenges that are still to be studied. Most of the works on unstructured big data quality are limited to specific cases of textual data; and DQD's to consider when dealing with unstructured Big Data analysis [33]. The authors identified important DQDs such as relevance, comparability, timeliness, accuracy, coherence, accessibility and ambiguousness related to specific Big data lifecycle phases. In [25], the authors identified some quality metrics to be used to evaluate unstructured data such as images usefulness and textual data readiness. An initial overview of quality assessment of UBD including quality indicators that must be used for unstructured data. These indicators are used in the traditional DQD and their scores are converted.

In [29], a definition of Unstructured data quality based on the similarity of input data to the data expected by its consumers, and to data representing the real world. They characterize DQD's to be used in these similarity process and propose measurable quality indicators to assess UBD quality.

In [34], the authors insists to have a characterization of quality in order to exploit Web data, these characterizations are materialized in data Trustworthiness and provenance. They also target some aspects of Big Data quality using examples of sensor data quality.

Other authors targeted social media data as unstructured big data for the purpose of quality evaluation. In [35], the authors

Authorized licensed use limited to: R V College of Engineering. Downloaded on June 29,2024 at 08:57:41 UTC from IEEE Xplore. Restrictions apply.

redefined DQD's and metrics to adapt to Big data context of unstructured data twitter feeds. They defined a set of metrics to evaluate the quality such metrics include readability, completeness and usefulness. Then implemented a real time system to evaluate the quality of twitter stream. In [36], Anne et al. introduced a new architectural solution to evaluate and manage the quality of social media data within the processing phase of the big data lifecycle. The objective was to improve business decision making by providing real-time customer insights from twitter feeds data using sentiment analysis for customer satisfaction [37].

To the best of our knowledge, a need of an assessment model for Unstructured Big Data Quality is of paramount importance. Most of the existing works are limited to a particular aspect of quality assessments of big data or to a specific type of unstructured data. Moreover, no quality management of such data has been proposed. In this paper, we present a tentative Model of Unstructured Big data Quality Assessment from data collection to quality reports generation.

IV. Unstructured Big Data Quality Assessment Model

To address the challenges of assessing quality of unstructured Big Data, we propose a quality assessment model that selects quality dimensions specific to each data type and evaluates its extracted features. Since unstructured data has no columnar values, we use a quantitative approach of data quality based on data contents. The model illustrated in Figure 4, has several components that the data goes through to achieve at the end a quality assessment report.

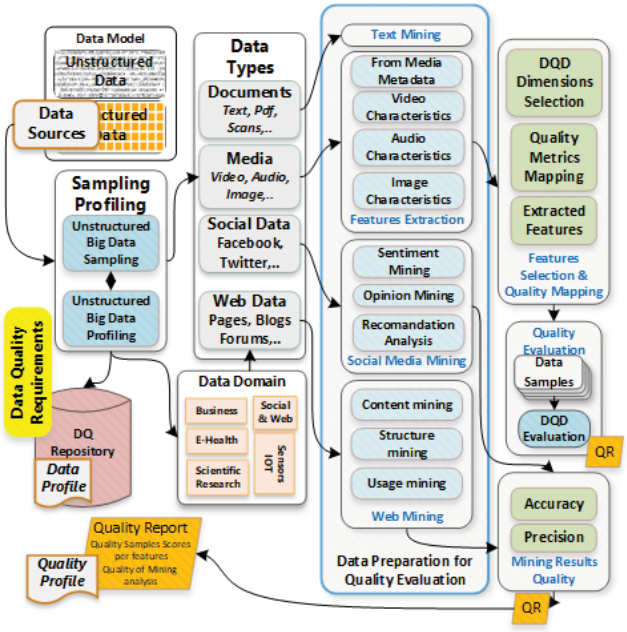

Fig. 4. Unstructured Big Data Quality Assessment Model

In the following, a description of each components and its interaction with the data flow being transmitted between is depicted in Figure 4.

1. **Quality Requirements:** the quality requirements are expressed in DQD acceptance scores, or a set of its indicators. Also, for some textual data a golden corpus or set is used as baseline for similarity analysis. For Media data like images, a set of image characteristics is defined with its related DQD's and Indicators. For example, High resolution image has higher number of pixels color levels.
2. **Data Sources**: represented as a collection of data sets or files, mixed or sorted by types, format or domain. This simple representation of data will help avoiding quality issues that may appear when auto-discovery of data types and domain is handled. This process is well known as data collection.
3. **Data Quality Repository:** stores and manages all related components of the data such as list of data domains, data types, and subtypes. DQD's, Quality Indicators, Metrics by data types. Data types features as in videos and pictures with their extraction functions. Predefined Mapping between DQD's, Features and metrics. It also stores the Data profile and Quality Report.
4. **Sampling and Profiling**: the data is sampled using BLB Bootstrap [38], then profiled to extract any metadata or useful information. After profiling, the data is identified by its type and domain if any. This will ease the next process of selecting which methodology to use for data quality based its data type. The sampling of media data is done differently for each type; for example, for Audio data, fragments are extracted and used to create a set of samples.
5. **Preparation Process for Quality Evaluation:** for textual data, text mining [39], [40] methodology is used to extract useful information about the textual data and assess its quality based on the extracted mining results. A definition of textual data quality indicators is of high importance to identify, create metrics that evaluate these indicators, combine and quantify them into a quality dimensions score. For videos data, feature extraction consists of discovering the format of the video first, then a sampling is applied to extract samples of video frames, afterwards extracting the set of features linked to these videos. Video characteristics consist of for example resolution (SD, HD, 4k), color saturation, and picture brightness. All these characteristics represent quality indicators of the video when the associated metrics are computed. The same scenarios apply for images and audio data.
6. **Feature Selection and Quality Mapping**: for each feature extracted from the videos, a DQD is selected with its related metric to evaluate its quality. The DQ repository contains a list of all common file types with all the predefined quality indicators and metrics.
7. **Quality Assessment:** after all the previous steps are completed, the assessment process runs the evaluation algorithm on sample data (set of videos frames, set of images, set of audio fragments) and quality report is built.

Authorized licensed use limited to: R V College of Engineering. Downloaded on June 29,2024 at 08:57:41 UTC from IEEE Xplore. Restrictions apply.

V. CONCLUSION

After the emergence of Big Data as a powerful way to extract insights from data where Unstructured data represented almost 80% from the overall data. Businesses and companies are highly interested to exploit this schema less data, while its quality is the key for its acceptance and usefulness. In this paper, we identified the key research challenges related to quality evaluation of Unstructured Big Data and we highlighted the importance of assessing the quality of such data. We surveyed and discussed the most comprehensive research initiatives in this area. We proposed our unstructured big data quality assessment model that emphasized the data exploitation and feature extraction activities that were conducted on Textual, Media and Web data. Finally, we debated the several techniques and methods to be used to assess the quality of unstructured data. We are currently implementing and experimenting different components of our model to deeply demonstrate how the quality of unstructured Big data activities can be integrated to the whole Big data Lifecycle.

REFERENCES

[1] R. Arsenault, "The Benefits of Utilizing Unstructured Data," *Aberdeen*, 01-Aug-2016. .

[2] J. Manyika *et al.*, "Big data: The next frontier for innovation, competition, and productivity," *McKinsey Glob. Inst.*, pp. 1–137, 2011.

[3] M. Chen, S. Mao, and Y. Liu, "Big Data: A Survey," *Mob. Netw. Appl.*, vol. 19, no. 2, pp. 171–209, 2014.

[4] C. P. Chen and C.-Y. Zhang, "Data-intensive applications, challenges, techniques and technologies: A survey on Big Data," *Inf. Sci.*, vol. 275, pp. 314–347, 2014.

[5] J. Wielki, "The Opportunities and Challenges Connected with Implementation of the Big Data Concept," in *Advances in ICT for Business, Industry and Public Sector*, M. Mach-Król, C. M. Olszak, and T. Pełech-Pilichowski, Eds. Springer International Publishing, 2015, pp. 171–189.

[6] I. A. T. Hashem, I. Yaqoob, N. B. Anuar, S. Mokhtar, A. Gani, and S. Ullah Khan, "The rise of 'big data' on cloud computing: Review and open research issues," *Inf. Syst.*, vol. 47, pp. 98–115, 2015.

[7] H. Hu, Y. Wen, T.-S. Chua, and X. Li, "Toward Scalable Systems for Big Data Analytics: A Technology Tutorial," *IEEE Access*, vol. 2, pp. 652–687, 2014.

[8] P. Kluegl, M. Toepfer, P.-D. Beck, G. Fette, and F. Puppe, "UIMA Ruta: Rapid development of rule-based information extraction applications," *Nat. Lang. Eng.*, vol. 22, no. 1, pp. 1–40, 2016.

[9] M. W. Berry and J. Kog, "Text Mining: Applications and Theory," p. 223.

[10] C. Rangu, S. Chatterjee, and S. R. Valluru, "Text Mining Approach for Product Quality Enhancement: (Improving Product Quality through Machine Learning)," in *2017 IEEE 7th International Advance Computing Conference (IACC)*, 2017, pp. 456–460.

[11] F. S. Gharehchopogh and Z. A. Khalifelu, "Analysis and evaluation of unstructured data: text mining versus natural language processing," in *2011 5th International Conference on Application of Information and Communication Technologies (AICT)*, 2011, pp. 1–4.

[12] B. Plale, "Big Data Opportunities and Challenges for IR, Text Mining and NLP," in *Proceedings of the 2013 International Workshop on Mining Unstructured Big Data Using Natural Language Processing*, New York, NY, USA, 2013, pp. 1–2.

[13] D. G. Chakraborty, "Analysis of Unstructured Data: Applications of Text Analytics and Sentiment Mining," p. 14.

[14] M. Salehan and D. J. Kim, "Predicting the performance of online consumer reviews: A sentiment mining approach to big data analytics," *Decis. Support Syst.*, vol. 81, pp. 30–40, 2016.

[15] A. Kaushik, A. Kaushik, and S. Naithani, "A Study on Sentiment Analysis: Methods and Tools."

[16] N. Tsirakis, V. Poulopoulos, P. Tsantilas, and I. Varlamis, "A platform for real-time opinion mining from social media and news streams."

[17] L. Dey and M. Haque, "Opinion mining from noisy text data," presented at the Proceedings of the second workshop on Analytics for noisy unstructured text data, 2008, pp. 83–90.

[18] M. Kang, J. Ahn, and K. Lee, "Opinion mining using ensemble text hidden Markov models for text classification," *Expert Syst. Appl.*, vol. 94, pp. 218–227, Mar. 2018.

[19] P. Oliveira, F. Rodrigues, and P. R. Henriques, "A Formal Definition of Data Quality Problems.," in *IQ*, 2005.

[20] M. Maier, A. Serebrenik, and I. T. P. Vanderfeesten, *Towards a Big Data Reference Architecture*. University of Eindhoven, 2013.

[21] M. Chen, M. Song, J. Han, and E. Haihong, "Survey on data quality," in *2012 World Congress on Information and Communication Technologies (WICT)*, 2012, pp. 1009–1013.

[22] F. Sidi, P. H. Shariat Panahy, L. S. Affendey, M. A. Jabar, H. Ibrahim, and A. Mustapha, "Data quality: A survey of data quality dimensions," in *2012 International Conference on Information Retrieval Knowledge Management (CAMP)*, 2012, pp. 300–304.

[23] P. Glowalla, P. Balazy, D. Basten, and A. Sunyaev, "Process-Driven Data Quality Management – An Application of the Combined Conceptual Life Cycle Model," in *2014 47th Hawaii International Conference on System Sciences (HICSS)*, 2014, pp. 4700–4709.

[24] C. Batini, C. Cappiello, C. Francalanci, and A. Maurino, "Methodologies for data quality assessment and improvement," *ACM Comput. Surv.*, vol. 41, no. 3, pp. 1–52, Jul. 2009.

[25] D. Firmani, M. Mecella, M. Scannapieco, and C. Batini, "On the Meaningfulness of 'Big Data Quality' (Invited Paper)," in *Data Science and Engineering*, Springer Berlin Heidelberg, 2015, pp. 1–15.

[26] A. McCallum, "Information extraction: Distilling structured data from unstructured text," *Queue*, vol. 3, no. 9, pp. 48–57, 2005.

[27] B. Carlo, B. Daniele, C. Federico, and G. Simone, "A Data Quality Methodology for Heterogeneous Data," *Int. J. Database Manag. Syst.*, vol. 3, no. 1, pp. 60–79, Feb. 2011.

[28] S. Malmasi, N. Hosomura, L.-S. Chang, C. J. Brown, S. Skentzos, and A. Turchin, "Extracting Healthcare Quality Information from Unstructured Data," *AMIA. Annu. Symp. Proc.*, vol. 2017, pp. 1243–1252, Apr. 2018.

[29] C. Kiefer, "Assessing the Quality of Unstructured Data: An Initial Overview."

[30] L. Cai and Y. Zhu, "The Challenges of Data Quality and Data Quality Assessment in the Big Data Era," *Data Sci. J.*, vol. 14, no. 0, May 2015.

[31] A. G. George and A. K. Prabavathy, "A Survey On Different Approaches Used In Image Quality Assessment," vol. 3, no. 2, p. 7, 2013.

[32] P. Ndjiki-Nya, M. Barrado, and T. Wiegand, "Efficient Full-Reference Assessment of Image and Video Quality," in *2007 IEEE International Conference on Image Processing*, 2007, vol. 2, pp. II-125-II–128.

[33] J. Maślankowski, "Big Data quality issues regarding unstructured data analysis. A case study," *Cover Title Page Des. ESENCJA Sp Z Oo*, p. 79, 2015.

[34] M. Scannapieco and L. Berti, "Quality of Web Data and Quality of Big Data: Open Problems," *Data Inf. Qual.*, pp. 421–449, 2016.

[35] F. Arolfo and A. Vaisman, "Data Quality in a Big Data Context," in *Advances in Databases and Information Systems*, vol. 11019, A. Benczúr, B. Thalheim, and T. Horváth, Eds. Cham: Springer International Publishing, 2018, pp. 159–172.

[36] A. Immonen, P. Paakkonen, and E. Ovaska, "Evaluating the Quality of Social Media Data in Big Data Architecture," *IEEE Access*, vol. 3, pp. 2028–2043, 2015.

[37] G. Gautam and D. Yadav, "Sentiment analysis of twitter data using machine learning approaches and semantic analysis," in *Contemporary computing (IC3), 2014 seventh international conference on*, 2014, pp. 437–442.

[38] I. Taleb, H. T. E. Kassabi, M. A. Serhani, R. Dssouli, and C. Bouhaddioui, "Big Data Quality: A Quality Dimensions Evaluation," in *2016 IEEE Conf(UIC/ATC/ScalCom/CBDCom/IoP/SmartWorld)*, 2016, pp. 759–765.

[39] N. Pavlopoulou, A. Abushwashi, F. Stahl, and V. Scibetta, "A Text Mining Framework for Big Data," *Expert Update*, vol. 17, no. 1, 2016.

[40] A. Hotho, A. Nürnberger, and G. Paa\s s, "A brief survey of text mining.," in *Ldv Forum*, 2005, vol. 20, pp. 19–62.

Authorized licensed use limited to: R V College of Engineering. Downloaded on June 29,2024 at 08:57:41 UTC from IEEE Xplore. Restrictions apply.