DIGIT.B4 – Big Data PoC

DIGIT 01 - Social Media

D02.02 Technological Architecture

everis Spain S.L.U

Table of contents

Table of tables

Table of figures

1 INTRODUCTION

The construction of the proof of concept is based on the definition of a technical architecture. The best way to define it is first to ensure that the PoC is built by following a solid methodological approach, and then to define a technical architecture supporting it.

The elements above are analysed in this document, which includes the following sections:

- Methodological Approach.
- Definition of the Technical Architecture.

2 METHODOLOGICAL APPROACH

From a methodological point of view, the main tasks that will be carried out are:

- Text-mining treatment: transform the content of the posts into a format that serves as an input for modelling algorithms
- Clustering: make homogenous groups of similar posts
- Classification model: obtain the rules to classify a document in one of the categories defined
- Draw conclusions from the classification of documents

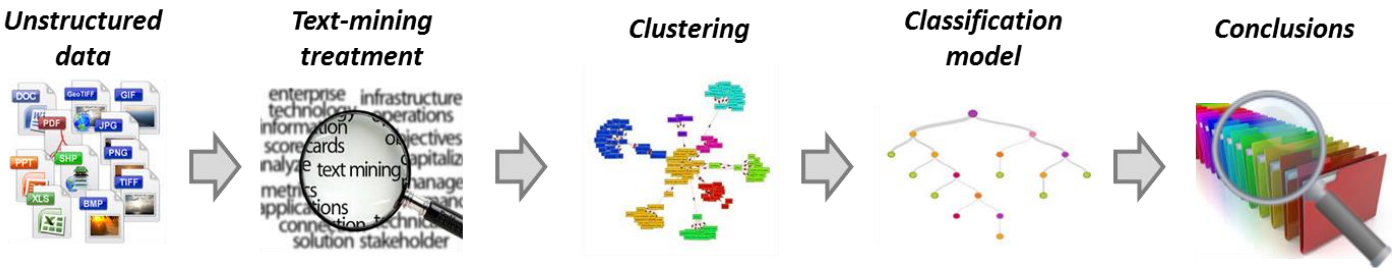

Figure 1 - Data analysis main flow

The activities are based on the CRISP-DM (Cross Industry Standard Process for Data Mining) methodology. This methodology is the most commonly used standard process model in both academic and industrial fields, and provides a solid framework for data mining projects.

The methodology displays a process in six phases which accounts for all the activities required to gather, classify, store and analyse the data. The CRISP-DM methodology has been chosen due to previous success cases in everis using it.

Phases	Business understanding	Data linguistic understanding	Data preparation	Modeling	Evaluation	Deployment
Objectives	Understand the project objectives and requirements from a business perspective, and then convert this knowledge into a text mining problem definition	Activities in order to get familiar with data, to identify data quality problems, to discover first insights into data, or to detect interesting subsets to form hypothesis	Activities to construct the final dataset. Correction of spelling errors and building dictionaries	Various modeling techniques are selected and applied, and their parameters are calibrated to optimal values	Evaluating and reviewing the quality of the model including whether the result	Transform the knowledge gained within the business process based on customer requirements
Activities	1. Understand the project objectives 2. Understand the current situation 3. Understand the text mining objectives	1. Load of data 2. Corpus construction 3. Data understanding 4. Filtering information	1. Parsing 2. Correction of spelling errors 3. Building dictionaries 4. Synonym and hyperonymic 5. Polysemy 6. Stemming	1. Building indexes 2. Term document matrix 3. Score of relevance 4. Modeling 1. Wordcloud 2. Term relation 3. Semantic network 4. LSA - Latent Semantic Analysis	1. Result evaluation	1. Deployment

Figure 2 - Methodology phases flow

2.1 Business understanding

The objective of this phase is to understand the project objectives and requirements from a business perspective, and then convert this knowledge into a text mining problem definition.

The business orientation of the problem is to obtain a better understanding of the main topics that are being treated to support the orientation of research funds.

The available sources will be analyzed to establish the best suited solution to the problem.

2.2 Data linguistic understanding

Once de data is loaded, a corpus (large collections of texts representing a sample of a variety presented a machine -readable form) is created to make analysis in order to understand the information contained in the posts received and ensure the data quality and the project viability.

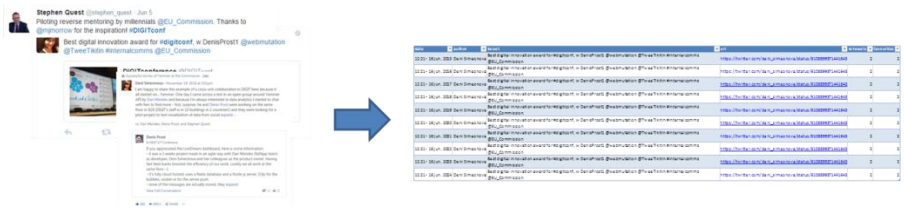

Figure 3 - Documents to corpus transformation

The main checks to make are:

- Completeness: the volume of documents and different words must be sufficient to allow reliable analysis.
- Quality: the words to be used for the analysis must be grammatically correct.
- Business point of view: the most frequent terms are analysed to validate quality from a business point of view.

Terms	Frequency
security	13.860
integrity	11.451
cloud	8.480
hosting	8.476
public administratio	8.207
citizen	7.851
mobile device	7.059
teleworking	6.587
big data	6.309
social media	5.657
search	5.446
BYOD	5.170
rationalization	5.150
smart machines	4.971
e-Government	4.668
outsourcing	4.279
paperless	3.792
...	...

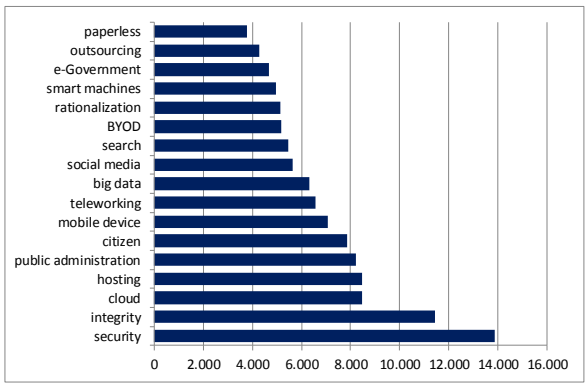

Figure 4 - Completeness and business standpoint analysis

2.3 Data preparation

In this phase, the necessary tasks are performed to transform the original text in a final dataset (set of keywords) that can be use by the algorithms in the modelling phase.

Corpus

Author	Date	Publication
Deni Simeonova	16/06/2015	Best digital innovation award for #digitconf, w DenisProst1 @webmutation @TweeTikitin #internalcomms @EU_Commission
Stephen Quest	05/06/2015	Piloting reverse mentoring by millennials @EU_Commission. Thanks to @mjmorrow for the inspiration! #DIGITconf
Dace Kalnina	19/11/2014	Thanks for having so cool 3d printer stand at #DIGITconf everybody loved it!

Algorithm input dataset

Publication
digital innovatiton award
Piloting reverse mentoring millenials inspiration
3d printer

Figure 5 - Data preparation process

The steps to transform the corpus are:

- Delete stop words: words that don't give any information and have no effect on how the rest of the words are related with each other.
- Correction of spelling errors.
- Stemming: the word stem is kept to unify conjugations.
- Create dictionaries: used to unify different terms that are expressing the same concepts
- Synonym and hyperonymic: used to unify terms that often appear together (these relations should be consistent from the perspective of business).

2.4 Modelling

During this phase, the statistics models are built to achieve the main objectives of the project and to assign a cluster and a classification to each document.

The first step is building the term document matrix (TDM) that will be the input for the LSA algorithm (will be explained later). This matrix contains the frequency of each term in every document.

Using this TDM, the score of relevance will determine which words are significant for clustering documents.

With these inputs, latent semantic analysis (LSA) is used to analyze links between a set of documents and the terms they contain by producing a set of concepts related to the documents.

Once the main clusters are defined, they are associatedto a concept classification (defined in the 2014 topics document).

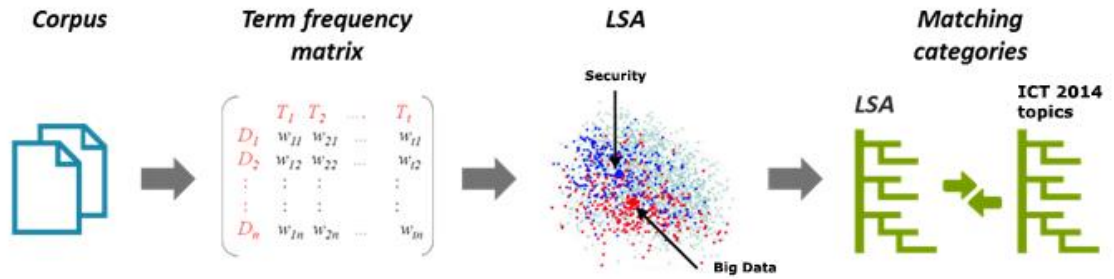

Figure 6 - Clustering process

This clustering is used as an input for a classification models. The result of executing a classification model is a set of rules that let us classify a new document based on the terms that appear in it.

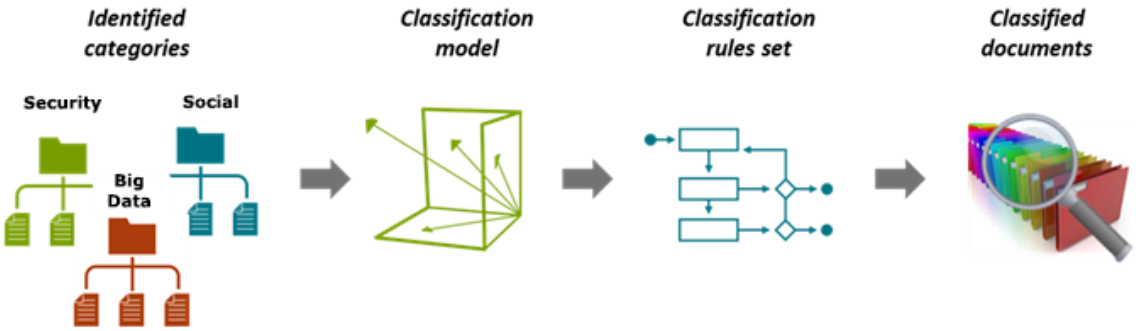

Figure 7 - Classification process

Before the execution of LSA and the classification model, other modelling tasks and analysis are made to know how to approach the main models:

- Wordcloud: to understand most frequent terms.

Figure 8 - Wordcloud

- Term relation: to identify related terms to define broader concepts that increase the effectiveness of classification and clustering algorithms

Security	
Term	% of occurrence
Authentication	35%
Confidentiality	27%
Integrity	13%
Protect	7%
...	...

Cloud and Hosting	
Term	% of occurrence
Hosting	50%
Cloud	34%
Migrating	10%
Public Cloud	5%
...	...

Mobile	
Term	% of occurrence
Mobility	73%
Mobile device	67%
Mobile computing	50%
IoT	43%
...	...

Figure 9 - Occurrence matrices

- Semantic network: to understand relationships between all terms in the different posts.

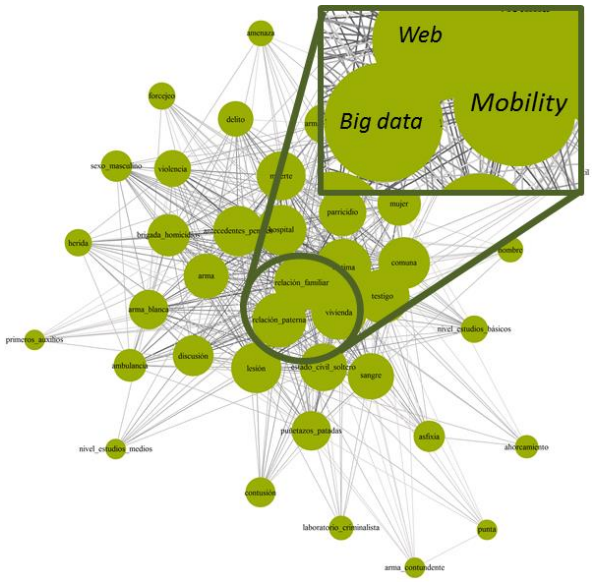

Figure 10 - Semantic network

2.5 Evaluation

In this phase some analysis are performed to ensure the quality of the generated models. It runs in parallel with the modeling phase.

All analysis and results of the execution of algorithms are tested to ensure statistical validity and check that they make sense from a business standpoint.

2.6 Deployment

Transform the results obtained after modelling in the analysis and conclusions required for the project.

Support results for presentation and visualization.

3 TECHNICAL ARCHITECTURE

The system architecture of the DIGIT.B4 – Big Data PoC (Proof of Concept) for the Social Media analysis will be based on the NTT Data – everis reference architecture.

The reference architecture is a collection of building blocks which consist of a series of technologies and methodologies.

NTT DATA and everis reference architecture is composed of 2 levels:

1. **Layer:** is a group of building blocks which have the same technology/methodology. There are 8 layers that are categorized for all technologies/methodologies important for the architecture.
2. **Building Blocks:** each one represents a technology/methodology. There are 38 building blocks in the architecture.

The blocks that will be used in the PoC for each layer have been highlighted.

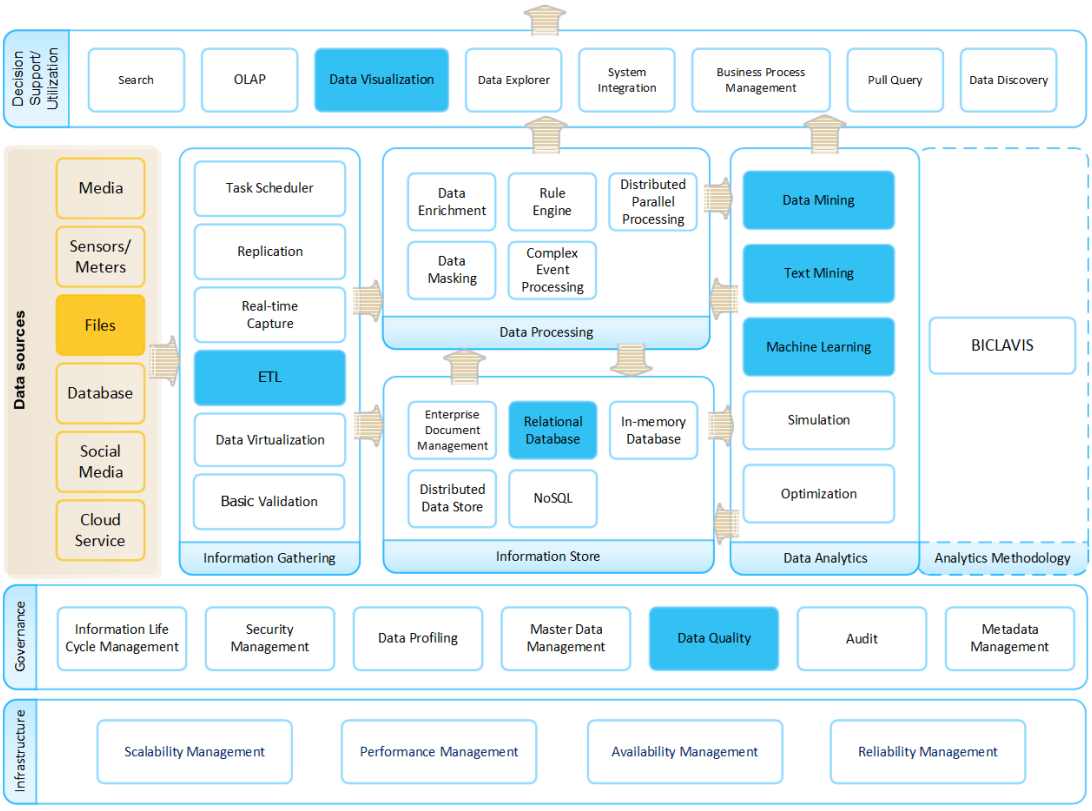

Figure 11 - System Architecture

The technologies chosen for each building block are listed, and the descriptions for the use that the PoC will make of each highlighted block are listed below:

3.1 Information gathering

ETL (Extract, Transform and Load)	
Building Block description	Extract, Transformation and Load (ETL) is an industry standard term used to represent the data movement and transformation processes.
Building Block PoC use	The files obtained from the data sources have to be transformed to other formats in order to be readable by the classification model.
Building Block technologies	R, python

Table 1- Information gathering

3.2 Information Store

Relational Data Base	
Building Block description	RDBMS (Relational Database Management System) is a software/middleware to manage relational data. It manages the data in the table that conforms to relational model.
Building Block PoC use	The gathered information will be stored in a structured manner.
Building Block technologies	Oracle Database, Oracle Exadata Machine, SAP IQ, SAP HANA, SQL Server, Parallel Data Warehouse, PureData, DB2, Teradata, Vertica, PIVOTAL, MySQL

Table 2 - Information Store

3.3 Data Analytics

Data Mining, Text Mining, Machine Learning	
Building Block description	**Data Mining**: Data mining refers to the statistical techniques and rule-based models that aim at discovering new knowledge and trends from a large data set. In contrast to machine learning, data mining emphasizes on the process of analysing data rather than the accuracy of the used models. **Text Mining**: Text mining can be defined as an intensive process in which a user interacts with a document collection over time by using a suite of analysis tools to extract knowledge. **Machine Learning**: Machine learning is a scientific discipline that explores the construction and study of algorithms that can learn from data. Such algorithms operate by building a model based on inputs that can be used to make predictions or decisions, rather than following only explicitly programmed instructions.
Building Block PoC use	The classification model and algorithms used for the PoC will use techniques from all of these blocks.
Building Block technologies	R, python

Table 3 - Data Analytics

3.4 Governance

Quality of Data	
Building Block description	The goal of a data quality exercise is to establish high-quality data and maintain it. The main tools/techniques that help to be successful in data health are related to the Master Data Management (MDM).
Building Block PoC use	The gathered data has to be treated to ensure its quality before sending it to the classification

	model.
Building Block technologies	R, python

Table 4 - Governance

3.5 Decision Support/Utilization

Visualization	
Building Block description	Data visualization means showing data in a graphical format. It involves the creation and study of the visual representation of data. The primary goal of data visualization is to communicate information clearly and efficiently to users via graphics, such as tables, charts and maps.
Building Block PoC use	The PoC results will be published in a web application where several charts and graphics will illustrate the outcome of the classification model.
Building Block technologies	R, JavaScript, HTML

Table 5 - Decision Support/Utilization