Mining Technical Knowledge from Texts

Engineering Management Methods & Natural Language Processing Techniques

Filippo Chiarello

2018-10-06

Contents

Preface

- Text Mining Techniques for Knowledge Extraction from Technical Documents

Part I

Introduction

Chapter 1

Problem

1.1 Big Data, Many Information, Less Knowledge?

Se l'ambiente informativo in cui noi esseri umani viviamo è radicalmente cambianto negli ultimi anni, con evidenti impatti su economia, tecnologia, cultura e società, l'impatto che ha avuto sulle aziende è ancora più forte(Rai et al., 2006; Jin et al., 2015b; Degryse, 2016; John Walker, 2014; O'Neil, 2016). Una persona infatti può raggiungere i suoi obiettivi anche senza il bisogno di gestire questa informazione (anche se con ormai alcune difficolà). Ciò non è vero per le aziende: per sopravviere ed essere competitive hanno la necessità di avere dei chiari metodi e strumenti per risolvere il problema della information overload (Levitin, 2014; Feng et al., 2015).

Per capire le sfide che devono affrontare le aziende oggi, dobbiamo inanzitutto svolgere un esercizio di immedesimazione nei loro confronti. Immaginiamo che le la quantità di informazione che potenzialmente ci interessa e che proviene dall'ambiente nel quale viviamo (le informazioni tattili, spaziali e sonore, cio che dicono le persone con le quali interagiamo, il contenuto dei documenti importanti per il lavoro che svogliamo ecc..) inizi velocemente ad aumentare. Per sopravvivere in un ambiente di questo genere, avremmo bisogno di sensi più potenti o forse di nuovi; potremmo inoltre aver necessità sistemi esterni che ci aiutino nel compito della processazione delle informazioni. Misurando la quantità di informazione digitale prodotta negli ultimi 10 anni, notiamo come questa situazione è simile a quella che le aziende stanno vivendo: si ritrovano in un universo digitale caotico ed in continua espansione. Questo universo digitale crescerà di un fattore di 300 dal 2005 al 2020, da 130 exabytes a 40,000 exabytes (40 zettabytes) (Gantz and Reinsel, 2012), come mostrato in figura 1.1.

Le nuove tecnologie digitali inoltre non stanno avendo un impatto solo fuori le aziende, ma anche dentro di esse (Lasi et al., 2014; Brettel et al., 2014; Rüßmann et al., 2015). Grazie ad industria 4.0 è infatti diventato possibile creare a poco prezzo un duale digitale dell'azienda, o in altri termini è economicamente fattibile oggi estrarre informazione da qualsiasi processo aziendale (Davies, 2015). Questo che pare essere un enorme vantaggio per le aziende, crea il problema di una information overload anche interna, che necessità di essere gestita proprio come quella esterna.

Un primo problema in tale contesto di pressione informativa crescente, è la difficoltà nel capire quali informazioni hanno valore per una azienda (Larose and Larose, 2014; Chemchem and Drias, 2015; Kasemsap, 2015) . Per valore qui si intende tutte le informazioni che possono aiutare l'azienda a costruire conoscenza in modo tale da perseguire la propia mission. Questo tipo di analisi richiede infatti sia una profonda comprensione delle tecnologie digitali, che di business acumen: non tutte le aziende sono in grado di avere al proprio interno (o comunque di procurarsi) queste skills (Hecklau et al., 2016; Davenport and Patil, 2012; Provost and Fawcett, 2013; Van der Aalst, 2014), considerando inoltre che il business acumen è fortemente dipendente dal settore nel quale si lavora. Come conseguenza è stato stimato che solo una piccola frazione dell'unverso digitale è stato esplorato con l'obiettivo di estrarre vantaggio competitivo (iView: IDC Analyze the future, 2012). La percenutale di dati ancora untapped è stimata essere del 25% ed è destinata a cresce al 33% entro

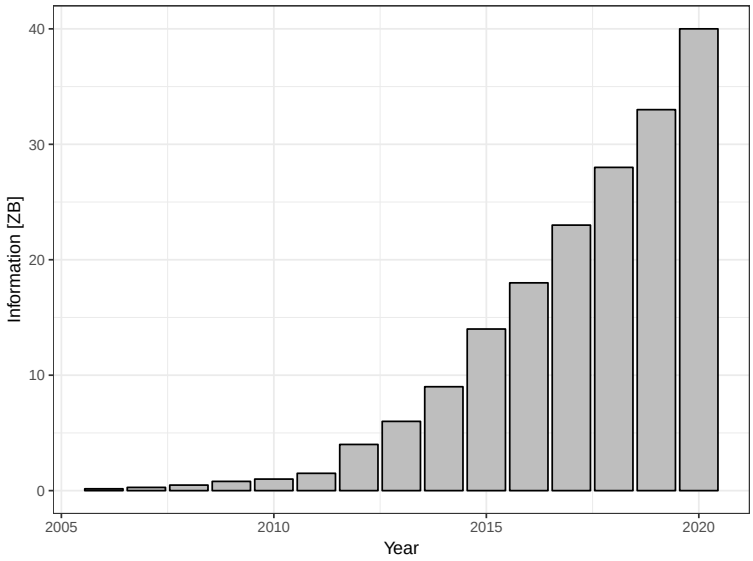

Figure 1.1: Hisogram of the eximated Zettabites of information produced from 2005 to 2020.

il 2020. Questo valore untapped può essere trovato in pattern nascosti nei social media, correlazioni fra studi scientifici, studi medici intersecati con studi sociologici, dall'analisi massima di documenti legislativi e così via.

Sempre in relazione alla difficoltà del reperimento di informazioni a valore, va considerato il fatto che sempre più spesso le informazioni ad alto impatto non sono più di tipo settoriale (tipiche di un solo knowledge field), ma nascono in contesti multidisciplinari. Questo le rende ancora più difficile da individuare. Tali informazioni anche se accessibili dagli esperti di un certo settore sono di difficile comprensione, poichè per definizione lontane dal settore stesso. A dimostrazione di tale fenomeno sta il recente aggiornamento della CPC (Cooperative Patent Classification) nella quale è stata introdotta la classe B33 e B33Y sulle tecnologie di addive manufacturing conosciuta anche come 3d-printing (Patent and amd European Patent Office, 2014), campo tecnologico ad alto contenuto multidisciplinare (scienze dei materiali, ingeria meccanica, informatica e robotica). Avere una nuova CPC significa un forte impatto sul patent officie, il quale deve foramre nuovi examiner per quella specifica CPC: questo è segno che i prcedenti examiner non avevano il range di competenze necessarie, e non è possibile a breve termine muoversi dall'interno di un dominio per conquistale.

In seconda istanza, note le potenziali fonti di informazione che possono portare valore, è necessario capire quali strumenti è possibile utilizzare per estrarre questo valore sotto forma di conoscenza(Hand, 2007; Mining, 2006) . Come mostrato nell'esempio precedente, una azienda si trova in una posizione di pericolo se la crescita di informazione sorpassa la propria capacità di processarla. In altri termini, l'aumento di informazione mostrato in figura 1.1 non implica un aumento di conoscenza, ma al contrario è probabile che contribuisca ad impedirne la crescita (Allen and Wilson, 2003; Herbig and Kramer, 1994). Un celebre esempio di tale fenomeno è il computer boom dei decenni 1970 e 1980, il quale implicò un declinio temporaneo nella produttività generale sia economica che scentifica (Solow, 1987). Tale fenomeno viene identificato con il nome di productivity paradox.

Dal punto di vista economico, evidenza di ciò è che i computer in quel periodo ebbero impatti su molti indicatori ma non su quelli di produttività economica (Robert Solow in 1987), e che gli stati uniti furono vittima di quattro recessioni tra il 1969 ed il 1982(of Economic Research, 2010).

Il progresso scientifico è più difficile da misurare di quello economico (Hirsch, 2005; Hauschildt, 1991; Van der Meulen and Rip, 2000; Ernø-Kjølhede and Hansson, 2011; Bornmann, 2013; Bornmann and Marx, 2014; Bornmann and Haunschild, 2017). Non è infatti chiaro che misure possano essere utilizzate per misurare quanto uno stato sia efficace nel creare conoscenza, nonostante la letteratura sterminata su tale tema. Una proxy largamente utilizzata per effettuare tale misura, è il numero di brevetti prodotti, letto in proporzione

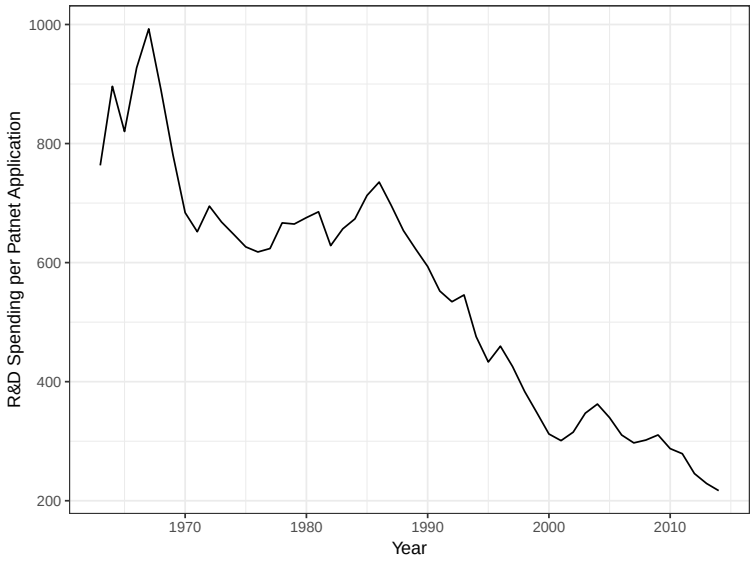

Figure 1.2: Trend in time (from 1964 to 2014) of the Research and Development Spending per Patent Application in dollars

rispetto a l'investimento in ricerca e sviluppo: se diventa meno costoso per le aziende innovare, questo suggerisce che le aziende stanno usando le informazioni che hanno a disposizione in maniera più efficiente e che sono in grado di trasformarle in conoscenza utilie all'innovazione. Per misurare tale costo, si può vedere quanto una nazione spende ogni anno per produrre in media un brevetto. Tale metrica ha ovviamente una serie di problematiche (fra le quali gli investimenti per brevettare non provengono solo da fonti pubbliche, un brevetto può far parte di una famiglia di brevetti e la United Stated Patent Office può processare brevetti non sviluppati in America) ma è interessante notare come nel 1960 gli stati uniti hanno speso $1000 (aggiustati per l'infalizione al 2015) per ogni patent application. Tale spesa inizialmente scende, poi sale anzichè scendere con l'arrivo dei computers con un picco di 735$ nel 1986 (Silver, 2012). La produttività ha nuovamente un crescita negli anni '90, quando ormai i computer erano diventati di largo utilizzo non solo per il business ma anche per applicazioni di tutti i giorni. La gobba presente intorno al decennio '80 nel grafico in figura 1.2 mostra come il sistema America è stato meno efficente nel produrre nuova conoscenza a seguito dell'introduzione dei computer nelle aziende.

Oggi, nell'era dei big data, le aziende (ma anche le università) hanno dunque bisogno di nuovi "sensi" ed "aiutanti" che le assistano nella estrazione e consolidazione della conoscenza. La presente tesi ha come obiettivo comprendere quali fonti di informazioni ad oggi contengono maggiore valore ancora non dischiuso e quali metodologie e strumenti posso essere utilizzati a tale scopo.

1.2 A New Challenge for Management Engineers

Tipicamente ci occupaimo di attività ad altà ripetitività. Ti porti dietro metodologie ingegneristiche applicate a sistemi inernti, andnano a operare in sistemi socio-tecnici. Hai fatto il tuo mestieri (ricerca operativa ecc..).

Negli ultimi anni però le aziende le attività a maggior valore aggiunto sono non ripetitive. R&S, Design, marketing, HR ecc.. e quindi gestione della conoscena. Su situazione che sembrano uniche il gestionale rischia di perdere rispetto al creativo. Come disciplina voglio presidiare queste aree: non ci occupaimo di casi unici, ma costruire modelli in grado di incorporare conoscenza per essere usati in questi.

La tesi ha l'obbiettivo di esploration and exploitation queste direzioni.

Chapter 2

Solutions

2.1 A Process for Knowledge Creation

Modello generico di come un sistema (uomo, macchina, azienda…) genera conoscenza. Immagine + spiegazione

Un modello più actionable: data science. Immagine + spiegazione

2.2 Text Mining

Istanza di data science e zona a maggior valore per estrazione conoscenza.

Importanza di definire documenti dai quali fare mining. Come si decide dove sta la conoscenza per l'azienda? Di quali documenti l'azienda è stakeholder?

2.3 Human & Machines

Datascience e text mining non fattibile da sole macchine…

Il problema non è sostituire domain knowledge. Idea vecchia ha fallito. E' insostibuibile perchè:

- Technology, interessa gli ingegneri
- Social Science, decision making

Perchè fallita: da una parte è andata avanti la knowledge rappresentation. E' impossibile rappresentare la conoscenza con regole, ma con altri strumenti si può rappresentare (bottom-up).

Inoltre ho text mining, capaità di processare testi. Parte di intelligenza artificiale. Questi fenomeni non sostituioscono l'esperto ma ne cambiano il modo di operare.

Si ha vantaggio su mitigazione bias se sistemi disegnati bene.

Il pericolo delle black boxes.

Oggi si integra. Vogliamo un esperto di dominio che faccia meglio il suo mestiere.

Abbiamo oggi più potenza e correzione errori.

Oltre ad efficienza e potenza nel correggere gli errori. Ora c'è anche la possibilità dio maggiore specificità. L'obiettivo è qiuindi poratre domain knowledge sia su technology sia ai decisori sociali.

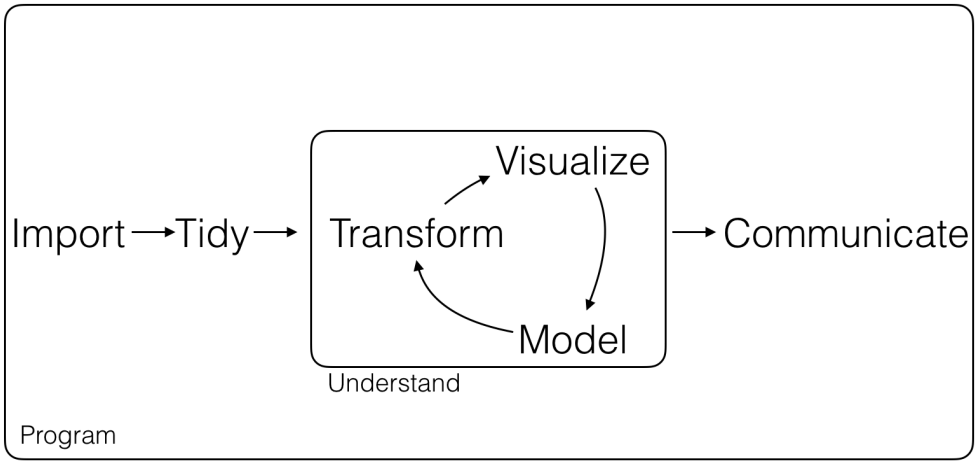

Figure 2.1: A general workflow for the process of data analysis. Readapted from Wickham (2016)

Chapter 3

Scope and Stakeholders

Algoritmi vecchi per domande nuove. Quale source per miglior precision noti gli algoritmi? Problema vincolato. Alcuni algoritmi.

Alcuni documenti in analisi, alcuni stakeholders

Research and Development, Design, Marketing, Human Resources. Policy makers.

Chapter 4

Structure and rationale

Part II

State of the Art

The analysis of technical documents require the design of processes that rely both on programming and Natural Language Processing techniques and on the understanding and knowledge of field experts. While the first techniques are codified and explicit, the second are sometimes implicit and always harder to systematize. In this section i treat these two groups of techniques in the same way to give to the reader a systematic literature review on these topics. For this reason the chapter of this part has the subsequent structure:

- At a first level there are two sections 5 and 6, reviewing respectively the processes of *programming and Natural Language Processing* and of *undestanding and knowldege of field experts application*;
- Section 5 has a subsection for each of the *phases* showed in figure 2.1. These subsections goes from 5.1 to 5.7;
- Each subsection from 5.1 to 5.7 contains the relative Natural Language Processing *task* that are relevant for the analysis of technical documents, for example Document Retrieval 5.2.1, Part-Of-Speech-Tagging; 5.4.6 or Named Entity Recognition 5.5.5.
- Each task subsection describes the relevant *techniques* to perform that task. I use the word techniques to include mainly algorithms and procedures but also more generic methods or frameworks;
- Since the second section 6 describes less systematic phases, task and techniques this section opens with a first subsection 5.8 that focuses on the studies of the problems of using expert knowledge in an analytic process and which are the techniques to convert this knowledge in a format that is usable in a Natural Language Processing workflow.
- Finally, always section 6 has a subsection for each of the anlyzed technical *documents*. These subsections goes from 6.1 to 6.4.

Chapter 5

Phases, Tasks, and Techniques

In this section I make a review of the most important techniques for Natural Language Processing in the context of technical documents analysis. The techniques (mainly algorithms) are grouped in phases (Import, Tidy, Transform, Model, Visualize, Communicate) showed in figure 2.1 and each phases is dived in the NLP tasks that are the most important for the analysis of technical documents. This standard process has been disclosed in the framework of the tidyverse (Wickham and Grolemund, 2016). The algorithms i reviewed in this section are summmarised in table tot, where the reader can see the relationship between tasks and techniques.

5.1 Program

Programming is a key activity to perform in order to effectively and efficiently perform text mining. It is not a phases per se because each phase is implemented trough programming. It is critical that an analysts has in mind the need of maximizing the probability that their analysis is reproducible, accurate, and collaborative. This goals can be reached only trough programming. The most used programming languages for text mining and natural language processing are R (R Development Core Team, 2008) and Python (Rossum, 1995). R and Python are both open-source programming languages with a large community of developers, and new libraries or tools are added continuously to their respective catalog. R is mainly used for statistical analysis and data science while Python is a more general purpose programming language. R has been developed by Academics and statisticians over two decades. R has now one of the richest ecosystems to perform data analysis and there are around 12000 packages available in CRAN (open-source repository of R). The rich variety of libraries makes R the first choice for statistical analysis. Another cutting-edge difference between R and the other statistical products is R-studio. RStudio is a free and open-source integrated development environment (IDE) for R. It includes a console, syntax-highlighting editor that supports direct code execution, as well as tools for plotting, history, debugging and workspace management. Finally, it is widely recognized the great performances that R has for data visualisation and communication. Python can pretty much do the same tasks as R: data wrangling, engineering, feature selection web scrapping, app and so on. Anyway, Python has great performances in the deployment and implementation of machine learning at a large-scale. Furthermore, Python codes are easier to maintain and more robust than R.

5.2 Import

The first activities to perform in a text mining pipeline is to find all the documents that contains useful information for the analysis and then import the corpus (the set of documents) in to the computer program. The present section is thus focused on techniques for document retrieval 5.2.1 and on the most popular documents digital formats 5.2.2.

5.2.1 Document Retrieval

Document retrieval is the process of matching a user query against a set of documents. A document retrieval system has two main tasks:

1- Find the documents that are relevant with respect to the user queries 2- Measure the relevance of the matching results

Building a query means to use field specific knowledge and logical rules to write a text string that is the composition of keywords and Boolean operators. The set of keywords (single words or phrases) is chosen in such a way that these are likely to be contained in the searched documents. Boolean operators can also be used to increment the performance of the query. The AND operator, for example is used to retrieve all the document that contains both of the terms at the left and the right of it, OR for document that contains at least one of the two words. Another important tool for making a good query are regular expressions. Regular expression (regexp) is a language for specifying text search strings, an algebraic notation for characterizing a set of strings. This language widely used in modern word processor and text processing tools. A regular expression search function will search through the corpus, returning all texts that match the pattern. For example, the Unix command-line tool grep takes a regular expression and returns every line of the input document that matches the expression. To evaluate the performance of a query is useful to understand the concepts of precision and recall.

Recall is the ratio of relevant results returned to all relevant results. Precision is the number of relevant results returned to the total number of results returned. Due to the ambiguities of natural language, full-text-search systems typically includes options like stop words to increase precision. Stop-words are words that filter all the document which contains them. On the other side, stemming to increase recall 5.4.3. The trade-off between precision and recall is simple: an increase in precision can lower overall recall, while an increase in recall lowers precision (Yuwono and Lee, 1996). Usually when a user performs a query, the main problem are false positives (the results that are returned by the systems but are not relevant to the user). False positives has a negative impact on the precision of the query. The retrieval of irrelevant documents is particularly strong for technical documents due to the inherent ambiguity of technical language. For this reason to understand and to use the rules of query building are fundamental to the technical document analysis, since without a good query is rare to have a good set of documents to analyze.

5.2.2 Documents Format

For the purpose of the present thesis documents are considered in a digital format, and there is no need to read it from a analogical source. From the computer science point of view, text is a human-readable sequence of characters and the words they form that can be encoded into computer-readable formats. There is no standard definition of a text file, though there are several common formats. The most common types of encoding are:

- ASCII, UTF-8: plain text formats
- .doc for Microsoft Word: Structural binary format developed by Microsoft (specifications available since 2008 under the Open Specification Promise)
- HTML (.html, .htm): open standard, ISO from 2000
- Office Open XML .docx: XML-based standard for office documents
- OpenDocument .odt: XML-based standard for office documents
- PDF: Open standard for document exchange. ISO standards include PDF/X (eXchange), PDF/A (Archive), PDF/E (Engineering), ISO 32000 (PDF), PDF/UA (Accessibility) and PDF/VT (Variable data and transactional printing). PDF is readable on almost every platform with free or open source readers. Open source PDF creators are also available.
- Scalable Vector Graphics (SVG): Graphics format primarily for vector-based images.

- TeX: Popular open-source typesetting program and format. First successful mathematical notation language.

For the R software there exist many packages that helps to import documents in several formats (Wickham et al., 2017).

5.3 Tidy

After that data are imported they have to be processed in such a way that it would be possible to perform the main task of data analysis (transformation, modelling and visualisation). This task of tidying data (usually referred to as data pre-processing) can be very time expensive, so it is important to have clear methods and techniques to perform this task.

Tidy data sets have structure and working with them is easy; they're easy to manipulate, model and visualize (Wickham et al., 2014). Tidy data sets main concept is to arrange data in a way that each variable is a column and each observation (or case) is a row. The characteristics of tidy data can be thus summarised as the points (Leek, 2015):

- Each variable you measure should be in one column
- Each different observation of that variable should be in a different row
- If you have multiple tables, they should include a column in the table that allows them to be linked

There main advantages of structuring the data in this way is that a consistent data structure make it easier to use the tools (programs) that work with it because they have an underlying uniformity. This lead to an advantage in reproducibility of code.

As stated before tidying data is not a trivial task, and applying this process to text is even harder for documents with respect to structured data (Silge and Robinson, 2016). On the other side, is clear that using tidy data principles can make many text mining tasks easier, more effective, and consistent with tools already in wide use . Treating text as data frames of individual words allows us to manipulate, summarize, and visualize the characteristics of text easily and integrate natural language processing into effective workflows already used.

Tidy text format is designed as being a table with one-token-per-row. A token unit of text that is meaningful for the analysis to be performed (for example letters, words, n-gram, sentences, or paragraphs). Tokenization is the process of splitting text into tokens. This one-token-per-row structure is in different from the ways documents are often stored in current analyses, mainly strings or document-term matrix. The term document matrix has each corpus word represented as a row with documents as columns. The document term matrix is the transposition of the TDM so each document is a row and each word is a column. The term document matrix or document term matrix is the foundation of bag of words text mining. The bag-of-words model is a simplifying representation of documents: a text is represented as the bag (multiset) of its words, disregarding grammar and even word order but keeping multiplicity (McTear et al., 2016).

5.4 Transform

Transforming in the context of Natural Language Processing is what in computational linguistic is called text normalization. Normalizing text means converting it to a more convenient, standard form. Most of the task of technical document analysis in fact relies on first separating out or tokenizing sentences and words, strip suffixes from the end of the word, determining the root of a word or transform the text using regular expressions.

5.4.1 Sentence Splitting

The analysis of technical documents require as first process, that the input text is segmented in sentences. Since documents do not encode this information in a non ambiguous manner (using dots) due to common abbreviations (e.g.: "Mr., Dr."), a sentence splitting process that does not rely only on a trivial *dot based* rule is required. This issue in the technical documents domain is even more problematic due to the presence of formulas, numbers, chemical entity names and bibliographic references. Furthermore, since sentence splitting is one of the first processes of an NLP pipeline, errors in this early stage are propagated in the following steps causing a strong decrease for what concerns their accuracy. One of the most advanced techniques are machine learning techniques: given a training corpus of properly segmented sentences and a learning algorithm, a statistical model is built. By reusing the statistical model, the sentence splitter is able to split sentences on texts not used in the training phase. ItalianNLP lab systems uses this approach (Dell'Orletta, 2009; Attardi and Dell'Orletta, 2009; Attardi et al., 2009). For this reason this algorithm is used for the most of the application presented in this Thesis.

5.4.2 Tokenization

Since documents are unstructured information, these has to be divided into linguistic units. The definition of linguistic units is non-trivial, and more advanced techniques can be used (such as n-gram extraction) but most of the times these are words, punctuation and numbers. English words are often separated from each other by white space, but white space is not always sufficient. Solving this problems and splitting words in well-defined tokens defined as tokenization. In most of the application described in the present Thesis, the tokenizer developed by the ItalianNLP lab was integrated (Dell'Orletta, 2009; Attardi and Dell'Orletta, 2009; Attardi et al., 2009). This tokenizer is regular expression based: each token must match one of the regular expression defined in a configuration file. Among the others, rules are defined to tokenize words, acronyms, numbers, dates and equations.

5.4.3 Stemming

Stemming is a simpler but cruder methodology for chopping off of affixes. The goal of stemming is reducing inflected (or sometimes derived) words to their word stem, base or root form. The stem of a word and its morphological root do not need to be identical; it is sufficient that related words map to the same stem, even if this stem is not a valid root. One of the most widely used stemming is the simple and efficient Porter algorithm (Porter, 1980).

5.4.4 Lemmatisation

Lemmatization is the task of determining the root of a words. The output allow to find that two words have the same root, despite their surface differences. For example, the verbs *am*, *are*, and *is* have the shared lemma *be*; the nouns *cat* and *cats* both have the lemma *cat*. Representing a word by its lemma is important for many natural language processing tasks. Lemmatisation in fact diminish the problem of sparsity of document-word matrix. Furthermore lemmatisaion is important for document retrieval 5.2.1 web search, since the goal is to find documents mentioning motors if the search is for motor. The most recent methods for lemmatization involve complete morphological parsing of the word (Hankamer, 1989).

5.4.5 Words importance metrics

Once that a document has been tokenized and the tokens has been transformed, an analyst usually wants to measure how important a word is to a document in a collection or corpus. Some of the metrics adopted are:

- *Term Frequency*: the number of times that a term occurs in document.
- *Boolean frequency*: 1 if the term occurs in the document and 0 otherwise;

- *Term frequency adjusted for document length*: is raw count normalized for the number of words contained in the document
- *Logarithmically scaled frequency*: is raw count normalized for the natural logarithm of one plus the number of words contained in the document
- *Inverse document frequency*: is the logarithmically scaled inverse fraction of the documents that contain the word, obtained by dividing the total number of documents by the number of documents containing the term, and then taking the logarithm of that quotient. It is a measure of how much information the word provides, that is, whether the term is common or rare across all documents.
- *Term frequency–Inverse document frequency*: the product between *term frequency and inverse document frequency*. A high weight in tf–idf is reached by a high term frequency (in the given document) and a low document frequency of the term in the whole collection of documents; the weights hence tend to filter out common terms.

5.4.6 Part-of-Speech Tagging

The part of speech plays an central role in technical document analysis since it provides very useful information concerning the morphological role of a word and its morphosyntactic context: for example, if a token is a determiner, the next token is a noun or an adjective with very high confidence. Part of speech tags are used for many information extraction tools such as named entity taggers (see section 5.5.5) in order to identify named entities. In typical named entity task these are people and locations since tokens representing named entities follow common morphological patterns (e.g. they start with a capital letter). For the application to technical documents, technical entities (like the possible failures of a manufact) becomes more relevant. In this context a correct part-of-speech tagger becomes even more important since morphosyntactical rules can not be used. In addition part of speech tags can be used to mitigate problems related to polysemy since words often have different meaning with respect to their part of speech (e.g. “track”, “guide”). This information is extremely valuable in patent analysis, and some patent tailored part-of-speech tagger has been designed (see section 6.1). The literature on pos-tagger is huge, and goes behind the scope of the present thesis to make a complete review. In most of the application presented in this work, was employed the ILC postagger (Attardi, 2006). This postagger uses a supervised training algorithm: given a set of features and a training corpus, the classifier creates a statistical model using the feature statistics extracted from the training corpus.

5.5 Model

The goal of a model is to provide a simple low-dimensional summary of a dataset (Wickham and Grolemund, 2016). Ideally, the model will capture patterns generated by the phenomenon of interest (true signals), and ignore random variations (noise). A good model at the same time is able to capture the weak signals that cab be easily confounded with noise. These information is particularly valuable in the context of technical document analysis, where great technical insight could come weak quasi-invisible signals. (James et al., 2013)

Probabilistic models are widely used in text mining nowadays, and applications range from topic modeling, language modeling, document classification and clustering to information extraction. The present section contains a review of the most used methods used to model textual information.

5.5.1 N-Grams

An n-gram is a sequence of N n-gram words: a 2-gram (or bigram) is a two-word sequence of words like “credit card”, “3d printing”, or ”printing machine”, and a 3-gram (or trigram) is a three-word sequence of words like “3d printing machine”. Statistical model can be used to extract the n-grams contained in a document. A first approach has the of predicting the next item in a sequence in the form of a $(n-1)$–order Markov model(Lafferty and Zhai, 2001). The algorithm begin with the task of computing P(w|h), the probability of a word w given a word h.The way to estimate this probability is using relative frequency counts. To

do that the algorithms count the number of times h is followed by the w. With a large enough corpus it is possible to build valuable models, able to extract n-grams (Bellegarda, 2004). While this method of estimating probabilities directly from counts works for many natural language applications, in many cases a huge dimension of the corpus does make the model useful, and this is particularly true for technical documents (Brants et al., 2012). This is because technical language has a strong ratio of evolution; as new artifact are invented, new chunks are created all the time, and has no sense to continuously count every word co-occurrence to update our model(Gibson et al., 1994). A more useful method for chunk extraction fro technical document uses part-of-speech-tagging and regular expression. Once a document is pos-taggerd each word is associated whit a particular part of speech: each sentence is represented as a sequence of part-of-spech. Once this representation is ready, it is possible to extract only certain sequences of part-of-speeches, the ones that whit an high level of confidence are n-grams.

5.5.2 Document Classification

Classification is a general process that has the goal of taking an object, extract features, and assign to the observation one of a set of discrete classes. This process is largely used for documents (Borko and Bernick, 1963) and there exist many methods for document classification (Aggarwal and Zhai, 2012).

Regardless of technological sector, most organizations today are facing the problem of overload of information. When it comes to classify huge amount of documents or to separate the useful documents from the irrelevant, document classification techniques can reduce the process cost and time.

The simplest method for classifying text is to use expert defined rules. These systems are called expert systems or knowledge engineering approach. Expert rule-based systems are programs that consist of rules in the IF form condition THEN action (if condition, then action). Given a series of facts, expert systems, thanks to the rules they are made of, manage to deduce new facts. The expert systems therefore differ from other similar programs, since, by referring to technologies developed according to artificial intelligence, they are always able to exhibit the logical steps that underlie their decisions: a purpose that, for example, is not feasible from the human mind or black box-systems. There are many type of documents for which expert based classifiers constitute a state-of-the-art system, or at least part of it. Anyway, rules can be useless in situations such as: - data change over time - the rules are too many and interrelated

Most systems of documents classification are instead done via supervised learning: a data set of input observations is available and each observation is associated with some correct output (training set). The goal of the algorithm is to build a static model able to learn how to map from a new observation (test set) to a correct output. The advantages of this approach over the knowledge engineering approach are a very good effectiveness, considerable savings in terms of expert labor power, and straightforward portability to different domains.

In the supervised document classification process, is used a training set of N documents that have each been typically hand-labeled with a class: (d1, c1),....,(dN, cN). I say typically, because other less expensive methods could be designed, as it will be shown for the task of Named Entity Recognition (another supervised learning task, that classifies words instead of documents 5.5.5). The goal of the supervised document classification task is to learn a statistical model capable of assign a new document d to its correct class c C. There exist a class of these classifier, probabilistic classifiers, that additionally will tell us the probability of the observation being in the class.

Many kinds of machine learning algorithms are used to build classifiers (Aggarwal and Zhai, 2012), such as:

- *Decision Tree Classifiers*: Decision tree documents classifier are systems that has as output a classification tree (Sebastiani, 2002). In this tree internal nodes are terms contained in the corpus under analysis, branches departing are labeled by the weight (see section 5.3) that the term has in the test document, and leafs are labeled by categories. There exists many methods to automatically learn trees from data. A tree can be build by splitting the data source into subsets based on an test feature. This process is repeated on each derived subset in a recursive manner called recursive partitioning. The recursion is completed when the subset at a node has all the same value of the target variable, or when splitting no longer adds value to the predictions.

- *Rule Based Classifiers*: Rule-based classifiers are systems in which the patterns which are most likely to be related to the different classes are extracted from a set of test documents. The set of rules corresponds to the left hand side to a word pattern, and the right-hand side to a class label. These rules are used for the purposes of classification. In its most general form, the left hand side of the rule is a Boolean condition, which is expressed in Disjunctive Normal Form (DNF). However, in most cases, the condition on the left hand side is much simpler and represents a set of terms, all of which must be present in the document for the condition to be satisfied (Yang et al., 2004).
- *Support Vector Machines (SVM) Classifiers*: SVM Classifiers attempt to partition the data space with the use of linear or non-linear delineations between the different classes. The main principle of SVM algorithm is to determine separators in the feature space which can best separate the different classes (Joachims, 1998; Manevitz and Yousef, 2001).
- *Baeysian Classifiers*: Bayesian classifiers build a probabilistic classifier based on modeling the underlying word features in different classes. The idea is then to classify documents using the posterior probability of the documents belonging to the different classes on the basis of the word presence in the documents (Pop, 2006).
- *Neural Netword Classifiers*: The basic unit in a neural network is a neuron. Each neuron receives a set of inputs, which are denoted by the vector *Xi*, which are the values of the feature vector for a certain instance. Each neuron is also associated with a set of weights, which are used in order to compute a function of its inputs. Neural Networks Classifier are able, thank to a process called learning phase, to adjust their weights in such a way that the function is able to effectively classify new instances. Neural networks are nowadays one of the best method for documents classification, and are used in a wide variety of applications (Manevitz and Yousef, 2007). Great performances has also been reached by deep neural networks, which are neural networks whit a large number o neurons arranged in multiple layers (Lai et al., 2015; Kim, 2014).

5.5.3 Sentiment Analsysis

Sentiment analysis techniques are algorithms able to measure from text, people's opinions and emotions toward events, topics, products and their attributes (Pang et al., 2008). For example, businesses (particularly marketeers) are interested in finding costumers opinions about their products and services.

Thanks to the growth of social media (forums, blogs and social networks), individuals and organizations are producing a huge quantity of their written opinion. This has make it possible to scholars to study this phenomena and to develop many different and effective sentiment analysis techniques (Liu and Zhang, 2012). In the past decade, a considerable amount of research has been done by scholars and there are also numerous commercial companies that provide opinion mining services. However, measuring sentiment in documents and distilling the information contained in them remains a challenging task because of the diversity of documents from which is possible to extract sentiment.

The approaches to perform sentiment analysis are many. Among all, the most interesting for technical documents analysis are:

- *Dictionary Base Approaches* : This approach has the aim of collecting words that are clues for positive or negative sentiment. In literature these words are called opinion words, opinion-bearing words or sentiment words. Examples of positive opinion words are: good, nice and amazing. Examples of negative opinion words are bad, poor, and terrible. Collectively, they are called the opinion lexicon. The most simple and widely used techniques to produce a dictionary of opinion words is based on bootstrapping using a small set of seed opinion words and an online dictionary such as WordNet (Miller, 1995). The works that used this approach (Hu and Liu, 2004; Kim and Hovy, 2004), adopts a process that consist in two phases: first collect set of opinion words manually, then grow this set by searching in the WordNet for their synonyms and antonyms. The process stops when no more new words are found. After that a manual inspection can be carried out to remove and/or correct errors. Scholars has developed several opinion lexicons (Ding et al., 2008; Baccianella et al., 2010; Hu and Liu, 2004; Philip et al., 1966; Wiebe et al., 1999) The lexicon based approach has the characteristic

of being strongly context specific. This is an advantage when the goal is to design a method able to extract sentiment in a specific context (Chiarello et al., 2017), but is a major shortcoming if the goal is to design a general purpose method.

- *Supervised Learning Approaches*: Sentiment analysis can be formulated as a document classification problem with three classes: positive, negative and neutral(Mullen and Collier, 2004). Training and test sets of documents are typically collected from product reviews, movies reviews or are created by scratch using manual annotation. Any learning algorithm can be applied to sentiment classification (naive Bayesian classification, and support vector machines (Prabowo and Thelwall, 2009)). The crucial phase for Supervised Learning sentiment analysis is the features presentation of the data. It was shown (Pang et al., 2002) that using uni-grams (a bag of individual words) as features in classification performed well with either naive Bayesian or SVM. Subsequent research used many more features and techniques in learning (Pang et al., 2008).

5.5.4 Text Clustering

The goal of clustering methods is to find groups of similar objects in the data thanks to the measure of a similarity function (Jain and Dubes, 1988; Kaufman and Rousseeuw, 2009). Clustering techniques has been widely applied in the text domain, where the objects of the clustering can be documents (at different level of granularity) or terms. In the context of technical documents analysis Clustering is especially useful documents retrieval (Anick and Vaithyanathan, 1997; Cutting et al., 1993). Clustering problems has been and are studied widely outside the text domain. Methods for clustering have been developed focusing on quantitative/non-textual data (Guha et al., 1998; Han et al., 2001; Zhang et al., 1996).

In the context of text analysis, the problem of clustering finds applicability for a number of tasks, such as Document Organization and Browsing (Cutting et al., 2017), Corpus Stigmatization using documents maps (Schütze and Silverstein, 1997) or word clusters (Baker and McCallum, 1998; Bekkerman et al., 2001). It is useful also to use a Soft clustering approach, that associates each document with multiple clusters with a given probability.

However, standard techniques for cluster analysis (k-means or hierarchical clustering) do not typically work well for clustering textual data in general or more specific technical documents. This is because of the unique characteristics of textual data which implies the design of specialized algorithms for the task.

The distinguishing characteristics of the text representation are the following (Aggarwal and Zhai, 2012):

- There is a problem of course of dimensionality. The dimensionality of the bug-of-words representation is very large and the underlying data is sparse. In other words, the lexicon from which the documents are drawn may be of the order of millions, but a given document may contain only a few hundred words.This problem is even more serious for technical documents in which the lexicon is even more large.
- The words are correlated with one another and thus the number of concepts (or principal components) in the data is much smaller than the feature space. This necessitates the careful design of algorithms which can account for word correlations in the clustering process.
- The number of words (or non-zero entries) in the different documents may vary widely. Therefore, it is important to normalize the document representations appropriately during the clustering task.

The problems of sparsity and high dimensionality necessitate the design of specific algorithms text processing. The topic has been heavily studied in the information retrieval literature where many techniques have been proposed (Ricardo and Berthier, 2011).

5.5.5 Named Entity Recognition

Named Entity Recognition is the task of identifying entity names like people, organizations, places, temporal expressions or numerical expressions. An example of an annotated sentence for a NER extraction system

tailored for user entity extraction from patents, is the following:

Traditionally, guitar players or players of other stringed instruments may perform in any of a number of various positions, from seated, with the stringed instrument supported on the leg of the performer, to standing or walking, with the stringed instrument suspended from a strap.

Methods and algorithms to deal with the entity extraction task are different, but the most effective are the ones based on supervised methods. Supervised methods tackle this task by extracting relevant statistics from an annotated corpus. These statistics are collected from the computation of features values, which are strong indicators for the identification of entities in the analyzed text. Features used in NLP for NER purposes are divided in two main categories: - Linguistically motivated features, such as n-gram of words (sequences of n words), lemma and part of speech - External resources features as, for example, external lists of entities that are candidates to be classified in the extraction process.

The annotation methods of a training corpus can be of two different kinds: human based, which is time expensive, but usually effective in the classification phase; automatically based, which can lead to annotation errors due to language ambiguity. For instance driver can be classified both as a user (the operator of a motor vehicle), or not a user (a program that determines how a computer will communicate with a peripheral device). Different training algorithms, such as Hidden Markov Models (Eddy, 1996a), Conditional Random Fields (CRF) (Lafferty et al., 2001a) Support Vector Machines (SVM) (Hearst et al., 1998b), or Bidirectional Long Short Term Mermory-CRF Neural Networks (Lample et al., 2016; Misawa et al., 2017,) are used to build a statistical model based on features that are extracted from the analyzed documents in the training phase.

5.5.6 Topic Modelling

Topic modeling is a form of dimension reduction that uses probabilistic models to find the co-occurrence patterns of terms that correspond to semantic topics in a collection of documents (Crain et al., 2012). To understand topic modelling it is useful to understand its differences with clustering 5.5.4 and the problem they both solves: the course of dimensionality. Both these techniques has in fact the goal of representing documents in such a way that they reveals their internal structure and interrelations. Clustering measures the similarity (or dissimilarity) between documents to place documents into groups. Representing each document by considering the belonging to a group, clustering induces a low-dimensional representation for documents. However, it is often difficult to characterize a cluster in terms of meaningful features because the clustering is independent of the document representation, given the computed similarity. Topic modeling integrates soft clustering (assigning each element to a cluster with a given probability and not with a Boolean variable) with dimension reduction. Each document is associated with a number of latent topics: a topic can be seed as both document clusters and compact group of words identified from a corpus. Each document is assigned to the topics with different weights: this feature can be seen both as the degree of membership in the clusters, as well as the coordinates of the document in the reduced dimension space. The result is an understandable representation of documents that is useful for analyzing the themes in documents. Latent Dirichlet allocation (LDA) is a particularly popular method for fitting a topic model (Blei et al., 2003). It treats each document as a mixture of topics, and each topic as a mixture of words. This allows documents to "overlap" each other in terms of content, rather than being separated into discrete groups, in a way that mirrors typical use of natural language.

5.6 Visualize

Traditionally, statistical training has focused primarily on mathematical derivations and proofs of statistical tests: the process of developing the outputs (the paper, the report, the dashboard, or other deliverable) is less frequently analyzed. This problem influences also text mining (Parker, 2017). One of the most studied problems of output production is data visualisation. Data visualisation involves the creation and study of the visual representation of data (Friendly and Denis, 2001). Data visualization uses statistical graphics, plots, information graphics and other tools to communicate information in a clearl and efficient way. The main

process of data visualisation is the visual encoding of numbers. Numerical data may be encoded in many ways, using a wide range of shapes: the main used are dots, lines, and bars (Wickham, 2016). The main goal of visualizations is to help users (students, researchers, companies and many others) analyze and reason about evidences hidden in data. It is possible thanks to the ability of visualisation to make complex data more accessible, understandable and usable. Tables are generally used where users will look up a specific measurement, while charts of various types are used to show patterns or relationships in the data for one or more variables.

Da visualisation has become in the last year a well enstablished discipline thanks to the increased amounts of data created by Internet activity and an expanding number of sensors in the environment are referred to as "big data" or Internet of things. It is important to undeline how the way this data is comunicated present ethical and analytical challenges for data visualization practitioner (Bikakis, 2018). The field of data science and practitioners called data scientists help address this challenge (Loukides, 2011).

Users of information displays are executing (consciously or not) particular analytical tasks such as making comparisons or determining causality (Tufte et al., 1990). The design principle of the information graphic should thus support the analytical task, showing the comparison or causality (Tufte, 2006).

Graphical displays and principles for effective graphical display is defined as the ability to communicate complex statistical and quantitative ideas with clarity, precision and efficiency (Mulrow, 2002). For this reason graphical displays should:

- show the data
- induce the viewer to think about the substance rather than about methodology, graphic design, the technology of graphic production or something else
- avoid distorting what the data has to say
- present many numbers in a small space
- make large data sets coherent
- encourage the eye to compare different pieces of data
- reveal the data at several levels of detail, from a broad overview to the fine structure
- serve a reasonably clear purpose: description, exploration, tabulation or decoration
- be closely integrated with the statistical and verbal descriptions of a data set

In litterature are identified eight types of quantitative messages that users may attempt to understand or communicate from a set of data and the associated graphs used to help communicate the message (Few, 2012):

- Time-series: A single variable is captured over a period of time, such as the unemployment rate over a 10-year period. A line chart may be used to demonstrate the trend.
- Ranking: Categorical subdivisions are ranked in ascending or descending order, such as a ranking of sales performance (the measure) by sales persons (the category, with each sales person a categorical subdivision) during a single period. A bar chart may be used to show the comparison across the sales persons.
- Part-to-whole: Categorical subdivisions are measured as a ratio to the whole (i.e., a percentage out of 100%). A pie chart or bar chart can show the comparison of ratios, such as the market share represented by competitors in a market.
- Deviation: Categorical subdivisions are compared against a reference, such as a comparison of actual vs. budget expenses for several departments of a business for a given time period. A bar chart can show comparison of the actual versus the reference amount.
- Frequency distribution: Shows the number of observations of a particular variable for given interval, such as the number of years in which the stock market return is between intervals such as 0-10%, 11-20%, etc. A histogram, a type of bar chart, may be used for this analysis. A boxplot helps visualize key statistics about the distribution, such as median, quartiles, outliers, etc.
- Correlation: Comparison between observations represented by two variables (X,Y) to determine if they tend to move in the same or opposite directions. For example, plotting unemployment (X) and inflation (Y) for a sample of months. A scatter plot is typically used for this message.
- Nominal comparison: Comparing categorical subdivisions in no particular order, such as the sales

volume by product code. A bar chart may be used for this comparison.
- Geographic or geospatial: Comparison of a variable across a map or layout, such as the unemployment rate by state or the number of persons on the various floors of a building. A cartogram is a typical graphic used.

Data visualisation practionires has to consider whether some or all of the messages and graphic types above are applicable to their task and audience. The process of trial and error to identify meaningful relationships and messages in the data is part of exploratory data analysis.

5.6.1 The Grammar of Graphics

Even if trial and error is and will remain an important part of data visualisation, some works as tried to give to data visualisation practitioners a well structured framework able to guide the process of data visualisation. Among the many frameworks, the most used is the *Grammar of Graphics* (Wilkinson, 2006) and its implementation (Wickham et al., 2008). The grammar of graphics is a coherent system for describing and building graphs. Like other kind of grammars, it describes to basic rules to use the element of data visualization with the goal of comunicating some content. The main concept in the grammar of graphics is that graphs are made by multiple layers. Layers are responsible for creating the objects that we perceive on the plot. A layer is composed of four parts:

- *Data and aesthetic mapping*: Data are independent from the other components: we can construct a graphic that can be applied to multiple datasets. Along with the data, we need a specification of which variables are mapped to which aesthetics.
- *Statistical transformation*: A statistical transformation transforms the data, typically by summarizing them in some manner.
- *Geometric object*: Geometric objects control the type of plot that is created. For example, using a point geom will create a scatterplot, whereas using a line geom will create a line plot. Geometric objects can be classified by their dimensionality.
- *Position adjustment*: Sometimes there exist th need to tweak the position of the geometric elements on the plot, when otherwise they would obscure each other. This is most common in bar plots, where we stack or dodge (place side-by-side) the bars to avoid overlaps.

Multyple layers togheter are used to create complex plots.

Togheter with the layer the designer can control the *scale*. A scale controls the mapping from data to aesthetic attributes, and one scale for each aesthetic property used in a layer is needed. Scales are common across layers to ensure a consistent mapping from data to aesthetics.

After the decision of the scale, the designer has ti decid the *coordinate system* for the layer. A coordinate system maps the position of objects onto the plane of the plot. Position is often specified by two coordinates (x, y), but could be any number of coordinates. The Cartesian coordinate system is the most common coordinate system for two dimensions, whereas polar coordinates and various map projections are used less frequently. For higher dimensions, we have parallel coordinates (a projective geometry), mosaic plots (a hierarchical coordinate system), and linear projections onto the plane. Coordinate systems affect all position variables simultaneously and differ from scales in that they also change the appearance of the geometric objects.

Finally, the last element of the grammar are *facets*. Faceting makes it easy to create small multiples of different subsets of an entire dataset. This is a powerful tool when investigating whether patterns are the same or different across conditions. The faceting specification describes which variables should be used to split up the data, and how they should be arranged.

5.7 Comunicate

The last task to perform in the process of knowledge extraction from technical documents is communications. If it means to comunicate the results of an anlysis inside a team or to the world, it doesn not matter how great an analysis is unless it is impossibile to explain it to others (Wickham and Grolemund, 2016). For the purposes of the present thesis, the focus is on the review of technical mechanics of communication especialli in the R (R Core Team, 2018) enviroment. One of the most important innovation for the task of communication in data science is R Markdown (Allaire et al., 2018). R Markdown provides an unified authoring framework for data science, combining code, results, and comments. R Markdown documents are fully reproducible and support dozens of output formats, like PDFs, Word files, slideshows, and more.

R Markdown files are designed to be used in three ways:

- For communicating to decision makers, who want to focus on the conclusions, not the code behind the analysis.
- For collaborating with other data scientists (including future you!), who are interested in both your conclusions, and how you reached them (i.e. the code).
- As an environment in which to do data science, as a modern day lab notebook where you can capture not only what you did, but also what you were thinking.

Togheter with reports (and usually contained in them) there are visualisation. Making graphics for communication follow all the rules and framework previously revised in section 5.6, but when a graph has to be used to comunicate to a wide audience there are some more rules to follow. The reason why this happen is that the audience likely do not share the background knowledge of the anlysit and do not be deeply invested in the data. To help others quickly build up a good mental model of the data, the analyst need to invest considerable effort in making plots as self-explanatory as possible. For this reason has been developed many tools to help data scientist to make effective comunication graphs(Wickham, 2016; Chang et al., 2017; Sievert et al., 2017; Pedersen, 2018; Bastian et al., 2009) .

5.8 Understand

The most difficult challenge in technology intelligence is not how to detect the large trends- they are visible anyway. It is, rather, how to detect weak signals, or information that initially appears with low frequency, in unrelated or unexpected regions of the technology landscape, and associated with large noise (Apreda et al. 2016). These signals escape from traditional statistical detection techniques, exactly because it is difficult to distinguish them from pure statistical noise. Metadata are not the appropriate source of data for detecting weak signals. As a matter of fact, they can be detected only by using a fine-grained domain knowledge structure, or using the full text of documents. As an example, classification-based clustering has been shown to be flawed because the patent class used is usually only the first one listed in patents, generating loss of granularity (Benner and Waldfogel, 2008; Aharonson and Schilling, 2016).

Hypothesis

postulation

5.8.1 Domain Expertise

(collins)

Sheela Jasanow

Taleb?

5.8.2 The problem of byases

Each site typically contains a huge volume of opinionated text that is not always easily deciphered in long forum postings and blogs. The average human reader will have difficulty identifying relevant sites and accurately summarizing the information and opinions contained in them. Moreover, it is also known that human analysis of text information is subject to considerable biases, e.g., people often pay greater attention to opinions that are consistent with their own preferences. People also have difficulty, owing to their mental and physical limitations, producing consistent results when the amount of information to be processed is large. Automated opinion mining and summarization systems are thus needed, as subjective biases and mental limitations can be overcome with an objective sentiment analysis system.

5.8.3 The Importance of Lexicons for Technical Documents Analysis

Chapter 6

Documents

In this section contains a review of the main classes of technical documents analyzed in the present work. The documents are Patents, Papers, Wikipedia and Social Media.

6.1 Patents

Nowadays patent data can be used for planning technological strategy (Ernst, 2003). The focus on the technological usefulness of patent data is certainly a great advantage, but this huge research area could hide other useful application for patents. For example, in (Jin et al., 2015a) the authors consider on one side patents as a source to collect information about technologies and products, and on the other side manuals, handbooks and market reports to collect market information. Since patents are only technological documents many potential patent reader (e.g. designers, marketers) could be taken aside. Despite this problem, some researchers (Bonino et al., 2010) affirm that there is an increasing variety of readers: not only technician and researchers but also marketers and designers who have grown an interest in patent analysis. Nevertheless, to our knowledge there are no researches that aim at facilitating information extraction for non-technological focused patent readers.

The bias that patent are only tech-oriented documents is due to two main reasons:

- Patents are produced to disclose and protect an invention, their content is mainly technical and legal.
- 80% of technical information is not available elsewhere (Terragno, 1979), so patents are one of the most comprehensive resources for technical analysis.

Focusing on the second point, our hypothesis is that also a fraction of all the other kinds of information (e.g. marketing and sociological information) is not contained elsewhere and it will appear in public documents (e.g. manual handbooks and market reports) in 6-18 months (Golzio, 2012).

Unfortunately there are four aspects reducing the non-tech readers' ability to analyze patents efficiently. First of all, an increasingly high number of patent filings generates a massive information overflow [Bergmann et al. (2008); secondly, analyzing patents takes a long time and requires skilled personnel (Liang and Tan, 2007); the quality of patent assessment process is decreasing (Burke and Reitzig, 2007; Philipp, 2006) because of the reduced assessment time available for patent examiners; finally, activities like patent hiding, proliferation and bombing, contribute to the generation of confusion and to the loss of time in research and analysis phases (Fantoni et al., 2013). These problems affect non-tech oriented patent readers as well as typical readers, even though the impact may be stronger on the firsts.

The main difference between typical and non-tech patent readers is the information they focus on. *Patent attorneys* and *Intellectual Property (IP) managers* are interested in reading patents for legal reasons to orient the IP direction. Analyzing patents is the core of their work, so they are experts in finding the information they need. Furthermore, they can spend most of their work-time on the activity. On the other hand, usually

marketers and designers (taken as example of non-tech oriented readers) search users' behavioral changes and needs, market trends, designers' vision, R&D trends and competitors' strategies. In addition, they rarely work with patents, so they do not know what and how to search. Lastly, they have short time to spend on the activity, and they waste most of this time understanding the legal and technical jargon used in patents.

Due to the large amount of information contained in patents and the growing interest to exploit this information, huge efforts have been devoted to the development of systems source to automatically extract different kind of information from such an enormous and valuable data.

Many techniques introduced in order to extract textual information from patents come from extensive research advances in the Natural Language Processing field (NLP). NLP is an area of research and artificial intelligence which aims at teaching computers to understand and manipulate natural language text in order to performs different tasks such as information extraction, machine translation and sentiment analysis.

The field of technology intelligence has become so large in recent years that several efforts to review and summarize the various approaches have been undertaken (Abbas et al., 2014). There are several possible ways to classify the approaches. For example a used classification distinguish between Visualization techniques (Patent networks, Clustering) and Text mining (NLP-based, Property-function, Rule-based, Semantic analysis, Neural networks). Another possible classification is:

- *Metadata approaches*: methods that uses sources of information embedded in patents, such as claims structure or bibliographic information
- *Keyword approaches*: methods able to produce vector representations of the analyzed documents. Computed vectors can be used for many applications such as patent retrieval by keyword, or patent similarity matching. Even though this approach can be used for several tasks, it is not suitable to catch semantic relationships between entities in sentences. Furthermore, these methods use a blacklist to remove noisy words (Blanchard, 2007) or use predefined lexicons (Chiarello et al., 2017). The right design of such list dramatically impacts the final output of the analysis (Lee et al., 2009a, Lee et al. (2015a), Montecchi et al. (2013))
- *Natural Language Processing approaches*: methods based on grammatical and syntactical structures extracted by natural language processing tools, such as Part-of-Speech taggers and syntactical parsers. Unlike the keyword based approach, these methods are able to capture the relationships between the entities mentioned in sentences (Yoon et al., 2013a; Park et al., 2011a, 2013a).

Each approach allows to capture different types of information from patents and build a knowledge base which can be exploited by patent analysis tools. For this reason the right approach to be chosen to develop a patent analysis system depends on the task to be solved, on the information to be analyzed and on the computational resources involved to solve the task. Choosing a good trade-off between these factors is a strict requirement in particular when analyzing big patent sets.

6.1.1 Metadata Approaches

Patent documents has the following metadata:

- Patent office
- Inventors
- Affiliation of the inventors
- Filling date
- Publication date
- Address of the affiliation of the inventor
- Patent classifications
- References
- Assignee
- Affiliation of the assignee
- Address of the affiliation of the assignee

Furthermore they contain text content:

- Title
- Abstract
- Keyword
- Summary page
- Drawing set
- Background of the invention
- Brief summary of the invention
- Brief description of the drawings
- Detailed description of the invention
- Claim set

This does not mean that metadata allow a unique identification. Disambiguation of metadata remains a challenge in most cases. Only recently documents have started to include a unique identifier, following the cooperation among the main producers and users. The unique identifiers refer to the publication (DOI, Digital Object Identifier) or the author (author ID). However, metadata are written and stored in a standardized way, so that it is possible to categorize them. Issues of disambiguation refer mainly to the identity of individual entities (e.g. distinguish between two authors with exactly the same name and surname) but not of categories (e.g. distinguish between the name of an author and the name of a university).

Metadata approaches for patent analysis exploit three types of information:

- bibliometric information
- patent structure information
- patent review process information.

In this approach are usually considered both patent and non-patent literature: for example patents with a high number of citations in papers usually indicate a strong correlation with the foundation of a technology.

On of the main problems addressed using metadata approaches is measure of the technology life cycle stage. The information about at which stage of maturity a technology is, is an important aspect taken into account by who decides to invest. Since the life cycle of a product is clearly related by patent grants evolution (Andersen, 1999), this lead research to investigate on patent indices that can be considered as appropriate life cycle stage indicators. The main effort has been directed in the identification of the three different technology life cycle stages: introduction, growth and maturity (Haupt et al., 2007). In this work, the author took into account that several studies have shown that a S-shape evolution of the number of patent applications or even a double-S-shape is typical. Consequently, the author defined the concept of patent activity index as an appropriate life cycle indicator only if its mean value differs significantly between the life cycle stages. The results of the work can be summarised as follow:

1. Backward literature citations increase significantly only at the transition from introduction to growth;
2. Backward patent citations increase significantly at both stage transitions;
3. The number of forward citations decreases significantly at the transition from introduction to growth;
4. The number of dependent claims is significantly higher at later technology life cycle stages than in earlier ones;
5. The number of priorities referred to in a patent application is significantly higher at later technology life cycle stages than in earlier ones;
6. Examination processes take longer in the phases of introduction and maturity than at the growth stage.

The main limit of these methods is the need of assumptions for what concerns the shapes of the stages curves.

For this reason furthuter works introduced an unsupervised method able to automatically detect the number of life cycle stages and the transition times of the technology of interest (Lee et al., 2016). Here, seven time series patent indicator were taken into account:

- patent activity which allows to model the evolution of a pattern. In particular increasing and decreasing patterns are considered a change for what concerns the research and development activity;
- the number of technology developers in the analyzed temporal series. It has been shown that a great number of competitor enters in the initial stages of a technology's life cycle, but this number lowers in

the matu- rity stage
- the number of different patent application areas in the considered temporal series. This is an important indicator since it has been shown that the number of technology application areas are small in the first stages of their life cycles and in- creases in the later life cycle stages
- the number of backward citations. It has been shown that patents with an high number of backward citations have less relevance with respect to the patents with a lower number of citations
- the number of forward citations which expresses the technological value of a patent in the analyzed temporal period
- the duration of examination processes as the average time between the filing and granting dates.
- the number of claims belonging to the patent. The more the number of claims reported by the patents, the higher the correlation with novelty and the financial value is.

Another widely adressed problem using metadata is citation analysis. Publications, patents, technical standards or clinical guidelines include a section in which other documents are cited. Citation analysis argues that including the reference to another document is the result of an intentional act, whose meaning may differ according to the type of document, but is nevertheless always worth of consideration (Moed, 2006). The analysis of citations, initially developed in scientometrics and bibliometrics, has migrated to technology intelligence, following the initial concept of patent bibliometrics (Narin, 1994). Patent (or firms, or inventors) that cite the same prior art are clustered together. Patent citation networks are then generated (Karki, 1997; Érdi et al., 2013). In fact, citations form a network structure, whose graph-theoretic properties can be interpreted in technology intelligence exercises (Lee and Kim, 2017). Patent citation networks have properties of small world (Cowan and Jonard, 2004) and their degree follows a power law distribution (Chen and Hicks, 2004). Patent citation analysis can be used to identify trajectory patterns and technology structure and paths, that is, knowledge flows among firms and among subsectors of an industry. In standard citation analysis all citations are considered equal. This counting approach can be criticized because "it relies on the assumption that patents are equally significant" (Gerken and Moehrle, 2012), which is in contrast with the empirical evidence on the large differences in patent value. This assumption is therefore removed in more advanced techniques in which the structure of citations from patents gets a qualification. It may be possible, however, that citations to other patents are strategically made by applicants, for example by citing their own patents or hiding other relevant citations). Backward citations may not be associated to technological novelty if they deliberately point only to the state of the art (Rost, 2011). Citations introduced by examiners are also another potential source of bias (Alcacer and Gittelman, 2006). Finally, it has been reported that the total number of citations increased over time, leading to "citation inflation" and the loss of value (Hall et al., 2001).

Derived from citation analysis, co-citation analysis argues that documents that are cited by the same documents should be considered part of the same cluster (Small, 2006; Small and Sweeney, 1985). A variant of this technique, called author co-citation analysis (White and Griffith, 1981), cluster documents that are cited by the same authors. Co-citation analysis can also be used to create classification systems of patents (Lai and Wu, 2005).

6.1.2 Keywords Approaches

In the keyword based approach each patent is represented as a vector where each component measure the importance of a specific keyword, like explained in section 5.4.5. The keywords to be taken into account depend on the patent set under analysis and on the goal of the task. Keywords can be extracted automatically using a text mining module, manually by experts or with hybrid methods where domain experts judge the relevance and the quality of the extracted keywords in order to limit the results to the most important keywords. Once keyword vectors are obtained, tasks such as patent similarity can be easily computed by using standard distance measures like cosine similarity. In addition, the keyword extraction allows to define more complex patent similarities measures (Moehrle, 2010) that can be exploited for the development of patent analysis tools (Lee et al., 2009b, 2015b) such as mappers or patent search engines. The main goal of these works is to developed systems for building keyword-based patent maps to be used for technology innovation activities. The system is composed of a text mining module, a patent mapping module and a

patent vacancies identification module. Once a specific technology field is taken into account for analysis and a related patent set is extracted, the modules of the system are sequentially executed. The text mining module automatically identifies relevant keywords in each patent of the considered patent set. Once all the keywords are extracted, only the ones with the highest relevance are selected for a further screening by domain experts. The final set of keywords resulting from the screening process is then considered for building the patent keyword vectors on the considered patent set. Specifically each component of the patent vector holds the frequency the corresponding keyword in the considered patent. Once all the keyword vectors are computed, the patent mapping module is executed to generate the patent map. The mapping is calculated by executing the Principal Component Analysis (PCA) algorithm on all the vectors. The PCA method allows to map n-dimensional vectors on a rectangular planar surface in order to generate the patent map. Intuitively this method allows to find the most meaningful 2 dimensional projection that filters out the correlated components of a n-dimensional vectors. The result of applying this method over the patent keyword vectors is a meaningful patent mapping, in which each patent is mapped over a 2-dimensional surface. Once the patent map is computed, a vacancy detection module is executed on the patent map. The vacancy detection module identifies sparse areas which can be considered good candidates for a research investigation. For each interesting vacancy, a list of related patents is obtained by selecting the ones which are located on the region boundaries. On the calculated list, a set of information for each patent is computed. This information is used to capture the importance of a patent in this patent list. Features considered strong indicators of the relevance of each patent are the number of citations [38] and the number of average citations by patents in the patent list. Finally, emerging and declining keywords are computed by taking into account the time series analysis of the considered keywords in the patent list. This allows to identify promising technology trends that can considered for further investigation.

The metadata and keyword approaches has a long tradition but suffers from several limitations. First, the initial query based on keywords is usually produced by human experts, either on an individual basis or organized as a panel. In practice, one of the best skills of research centers or consultancies specialised in technology intelligence has been, in the past, the ability to mobilize high level experts on an international basis in order to produce well crafted query lists. Unfortunately these lists, even if they are produced following elicitation procedures that respect state of the art recommendations in social sciences, are inevitably biased. Experts are extremely good in their field, but are not better that others if they have to evaluate matters that are outside their domain (Burgman, 2015). To the extent that emerging technologies are complex and fast evolving technologies, it is likely that experts have a narrow, or biased, perception of the dynamics. Experts tend to keep their existing R&D areas in mind, have personal and organizational inclinations, are subject to halo effects in favor of well known institutions or solutions, and may follow different criteria for selecting promising technologies (Kim and Bae, 2017). It has been shown that little differences in the wording of queries, or on the time window, may end up in completely different sets of documents, leading the analysis in different directions (Bassecoulard et al., 2007). In addition to these authors, several studies in recent years have called the attention to the risk that initial differences in the delineation process generate non-comparable descriptions of technologies (Mogoutov and Kahane, 2007; Youtie et al., 2008; Ghazinoory et al., 2013). Following this line of concern,methodologies to update the keyword structure in an iterative, or evolutionary way has been proposed (Mogoutov and Kahane, 2007). Second, it has been shown that when experts are asked to decide on relatedness measures (e.g. synonyms, hypernims or hyponims), they do not apply systematic rules (Tseng et al., 2007; Noh et al., 2015). Third, the query list is static. Once defined, it is used to extract documents from large corpora, which are then processed. In dynamic technologies, it is likely that the pace of technological changes exceeds the speed of updating of the query lists. It is difficult to convene panels of experts repeatedly, also because of the large costs incurred in expert selection and management (Tseng et al., 2007). As an example, with the advent of nanotechnology it was felt the need to introduce a new patent sub-class. The sub-class B82B was introduced in year 2000, but it did not incorporate the previous patents, so that a comparison across time is not feasible. A new sub-class, B82Y, was introduced in 2011 (Kreuchauff and Korzinov, 2017).

6.1.3 Natural Language Processing approaches

The impressive advancements of computational linguistics in the last two decades have made it possible to carry out analysis on the full content, not only the metadata, of large collections of texts. In text mining patterns are extracted from unstructured collection of documents, while in the metadata approach the patterns are extracted from structured documents or databases. This has opened the way to the "full text based scientometrics" (Boyack et al., 2013) and has created the conditions for the convergence between the citationist approach illustrated above, and the lexical approach. Text mining techniques have then been applied to the corpus of patent texts, with a number of extremely powerful results (Tseng et al., 2007; Joung and Kim, 2017; Kreuchauff and Korzinov, 2017; Ozcan and Islam, 2017; Yoon and Kim, 2012). In turn, text mining can be applied for the search of specific words (or combination thereof) or in the search for patterns that are not defined ex ante. In the former case the most used techniques are combination of keywords, correspondence analysis or category specific terms. These approaches expand the search over the full text of patents but preserve the limitations of keyword-based search. On the contrary, the search for patterns is the object of the most largely used technique, namely topic modelling. Pattern recognition in patent texts is "still in its infancy" (Madani and Weber, 2016) but its applications are growing rapidly. A useful review of NLP techniques in patent analysis (Madani and Weber, 2016) identifies:

- the statistical approach that uses the Term Frequency-Inverted Document Frequency (TF-IDF) method to detect regularities
- the semantic approach uses SAO (Subject-Action-Object) and property-function structures in order to attribute meaning to the texts
- the corpus approach adopts ontology-based techniques.

In turn, all these three information retrieval approaches can be extended by using pattern recognition techniques, that are keyword-, patent- or concept-based.

Text mining has several limitations : it cannot consider synonyms and the co-occurrence of keywords, while the inclusion of compound words and n-gram expressions requires large computational power. In addition, in the case of patents, claims are written in "arcane legalese" in order to hide critical elements and confound potential competitors. The challenge here is how to maximize the substantive knowledge that can be generated by automatic processing of the full text. It has been remarked since long time that a promising direction for research into technology intelligence and foresight lies in the combination of methods. This recommendation requires the combination between domain-knowledge and powerful computational approaches. It is this combination that holds the best promise to generate methods for the identification of emerging technologies, and more generally, for technology intelligence, that are able to identify high-granularity information producing weak signals, that is, to distinguish accurately the signal from the noise in turbulent and dynamic technological landscapes.

By exploiting the information obtained by these steps, several information extraction tasks can be solved by other NLP tools such as: - Term extraction: the task of automatically extract relevant terms from a given corpus. Part of Speech tags are typically used by term extractors to narrow the terms search to a predefined term structure; - Named entity recognition: the task of automatically identify and classify named entities in text such as persons, organizations and locations. Named entity recognizers usually use Part of Speech tags in order to disambiguate the morphosyntactic role of tokens in a phrase, improving the performance of the extraction; - Relation extraction: the task of automatically build relations among entities in the analyzed text. In this context entities can be named entities or extracted terms. In addition, the syntactic role of the entities can be exploited to better categorize the relation type (e.g.: subject, object).

Technical domain language, as other linguistic domains, suffers from linguistic ambiguities. For instance the word "support" can have two totally different meanings when used as a noun or as a verb. By using part of speech taggers which are able to disambiguate the morphological role of each word in a sentence, more precise information extractions are possible and can be used in several applications (e.g. patent search engines). In addition part of speech taggers allow to perform textual lemmatization, which can further improve the performances of automatic patent analysis tools. Another key NLP tool used by several automatic patent analysis systems are syntactic parsers: by identifying the syntactic role of each word in document sentences, several patent analysis applications are possible.

The most well established system for patent analysis using NLP techniques is the extraction of the Subject-Action-Object (SAO) structures, which is also a common use of syntactic parsers in automatic patent analysis tools. Each SAO structure represents the subject (S), the action (A) and the object (O) in a patent sentence (Yoon and Kim, 2011b). By automatically extracting SAO structures from patents, relationships between key technological components can be easily represented (Yoon et al., 2013b; Choi et al., 2011; Park et al., 2011b).

Another techniques that is growing in patent literature analysis is Named Entity Recognition (for further details see section 5.5.5). The Named Entity Recognition (NER) is the task of identifying entity names like people, organizations, places, temporal expressions or numerical expressions.

Entity extraction tools used in patent analysis are largely based on NLP tools which can be applied to the analyzed text to extract entities that are important for the extraction purpose. For example, in the chemical field relevant entities are chemical components, proteins or product names. For the latter cases, adaptations of Named Entity Recognizers (NER) are commonly used for this task.

Methods and algorithms to deal with the entity extraction task are different, but the most effective are based on supervised methods. Supervised methods tackle this task by extracting relevant statistics from an annotated corpus. These statistics are collected from the computation of features values, which are strong indicators of the identification of entities in the analyzed text. Features used in NLP based entity recognition systems, are divided in two main categories:

- linguistically motivated features, such as n-grams of words, lemma and part of speech;
- external resources features as, for example, external lists of entities that are candidates to be classified in the extraction process.

The annotation methods of a training corpus can be of two different kinds: (a) human based, which is time expensive, but usually effective in the classification phase; (b) automatically based, which can lead to annotation errors due to language ambiguity. As an example *crack* can be classified both as a drawback (a fracture), or not drawback (short for crack cocaine). Different training algorithms, such as Hidden Markov Models (Eddy, 1996b), Neural Networks (Haykin and Network, 2004), Conditional Random Fields (Lafferty et al., 2001b) or Support Vector Machines (Hearst et al., 1998a), are used to build a statistical model based on the features that are extracted from the analyzed documents in the training phase. The same statistical model is later used in classification of unseen documents.

For what concerns the extraction of specific entities in patents, a major interest both in academia and commercial organizations has raised in the latest years, with the main aim of improving the accuracy of domain specific patent retrieval systems (Krallinger et al., 2015). In (Lee and Kang, 2014) the authors proposed a machine learning based patent NER system that identifies key terms in patent documents and recognizes products, services and technology names in patent summaries and claims. In this work a study was conducted to identify the most relevant features for this classification task and by using lexical features like word uni-grams, word bi-grams and word trig-rams, their NER system reached an F1 score (the harmonic mean of precision and recall) of 65.4%. The authors compared their NER tagging system resulting from the optimal feature selection method, with the human tagged corpus, showing that the kappa coefficient was 0.67. This result was better than the kappa coefficient between two human taggers (0.60).

Other entity extraction systems for the patent domain were proposed for the CHEMDNER (chemical compounds and drug names recognition) community challenge (Krallinger et al., 2015). The main aim of the organizers was to promote the development of novel, competitive and accessible chemical text mining systems. The best results were obtained by the *tmChem* system (Leaman et al., 2015), achieving a 0.8739 f-measure score. The authors proposed an ensemble system composed of two Conditional Random Fields based classifiers, each one using hard feature engineering such as lemmatization, stemming, lexical and morphological features. In addition, external lists of entities were exploited to recognize whether a token matched the name of a chemical symbol or element, each one used to compute features to be added in the final statistical model.

The described entity tagging systems have very good performances mainly for two reasons: firstly, chemical entity names (such as molecular formulas) have very common orthographic patterns; secondly, these entities surrounding contexts are very similar. In more generic cases, these two features can not be exploited for entity extraction from patents, since different words have totally different surrounding contexts. Another important

key factor concerning the high performances of the described systems is that many external resources, such as lists of chemicals or product names, are available: this external knowledge can not be fully exploited in generic system.

6.2 Papers

6.3 Wikipedia

The use of Wikipedia as source of knowledge started more than a decade ago and has been validated repeatedly in a variety of text mining applications (text annotation, categorization, indexing, clustering, searching (Milne and Witten, 2008)). In addition to the large and growing size in terms of number of articles, the structure of Wikipedia has a number of useful features that make it a good candidate for text mining applications. First, Wikipedia pages are considered reliable in many knowledge fields, including the ones more interesting for technical analysis, i.e. engineering and computer science (Xu et al., 2015). The pages are regularly and systematically updated by a large global community of contributors, which includes many scientific and industrial authorities in the field. The use of Wikipedia as knowledge source for computerized text mining tools is established in the literature (Ferragina and Scaiella, 2012). In addition, it is powerful in disambiguation of terms, particularly through the use of redirect pages and disambiguation pages. This means that it can be used for detection and disambiguation of named entities (Bunescu and Paşca, 2006). Second, the pages include links to other pages motivated by clear reasons on content. There are many links between Wikipedia pages, which are clues for semantic relations. This makes Wikipedia a densely connected structure, creating a classical small world effect: according to an often cited estimate, it takes on average 4.5 clicks to reach an article from any other article (Dolan, 2008). Unfortunately it is not possible to disentangle the kind of semantic relation, introducing a distinction between equivalent relations (synonymy), hierarchical relations (hyponymy/ hyperonimy) and associative relations, but this limitation is not relevant for our applications. Third, it makes use of categories which do not have a hierarchical structure, but a tree-like structure. Fourth, it has the ability to evolve quickly (Lih, 2004), particularly after the development of systems such as Wikify (Mihalcea and Csomai, 2007; Cheng and Roth, 2013). Wikipedia has by design a dynamic structure, since it is constantly growing in the number of entries and changing in their content, when this is needed due to the advancements of knowledge (Ponzetto and Strube, 2007). Furthermore the new terms that appear on Wikipedia thanks to comprehensive contributions by volunteers around the world, cannot be found in other linguistic corpora, such as WordNet Miller, 1995. Indeed, Wikipedia is the expression of a large international community, that is, of a "real community agreement" (Bizer et al., 2009) or "community consensus" (Hepp et al., 2007), guaranteed by permanent collective monitoring of the quality and rigor of the entries (Bryant et al., 2005). Finally, Wikipedia is free-content and multilingual. This make it possible to freely collect the information contained in the web pages and allows the possibility for future developments of the dictionary in other languages. In our opinion multilanguage is an interesting feature for the dictionary, due to the fact that Industry 4.0 is a worldwide phenomena.

These properties make Wikipedia the ideal candidate for the goal of extracting technical knowledge from texts. Technical fields are in fact comprehensive, dynamically updated, and, as far as possible, expert-independent. In particular, Wikipedia entries allow an endogenous measurement of semantic relatedness. This is an exceedingly important property for technical analysis: technologies can be mapped and can be defined as included in the perimeter of a knowledge field if and only it exhibits relatedness with other technologies already included in the perimeter. The inclusion of new technologies is therefore not dependent on experts' subjective views, but is endogenously generated by the technological community that writes the articles for the encyclopedia and includes hyperlinks in the text of newly added pages.

6.4 Social Media

Nowadays, more than ever before, companies, governments, and researchers can gather and access data about people on a massive scale. Monitoring public opinion is increasingly made possible thanks to the rise of Social Media. These ones are computer-mediated technologies that facilitate the creation and sharing of information, ideas, career interests and other forms of expression with friends, families, co-workers, and other users, via virtual communities and networks. There are many different Social Media platforms, each of which targets a different aspect of what users want or need: e.g., LinkedIn targets professional networking activities, Facebook provides a mean of connecting friends and family, and Twitter provides a platform from which to quickly broadcast thoughts and ideas. These platforms are incredibly popular: as of February 2017, Facebook sees an average of 1,871 billion active users, with 76% of them that logging in every day (Tuten and Solomon, 2017).

Being so widely used, Social Media platforms generate huge amount of data. In 2013 users were posting an average of over 500 million tweets every day (Krikorian, 2013). Social Media are not constrained by national, cultural, and linguistic boundaries differently from traditional data sources and records of human activities, such as newspapers and broadcast media. Moreover, traditional media requires time to compile relevant information for publication, while Social Media data is generated in real-time as events take place.

Virtually anyone who wishes to use all this information could collect and mine it. In 2009, the United States Geological Survey (USGS) began investigating the possibility of using SM data to detect earthquakes in real time (Ellis, 2015). Information about an earthquake spreads faster on Social Media than the earthquake itself can spread through the crust of the Earth (Konkel, 2013). Similarly, interesting work in Social Media forecasting also exists: EMBERS is a currently deployed system for monitoring civil unrests and forecasting events such as riots and protests (Ramakrishnan et al., 2014). Using a combination of Social Media and publicly-available, non-SM, researchers are able to predict not just when and where a protest will take place, but also why a protest may occur. These encouraging results have stocked the interest of researchers toward the possibilities opened by Social Media data, although some unanswered questions remain. If Social Media is useful for detecting real-time events, can it be used to make predictions about the future? What limitations does forecasting with Social Media data face? What methods lead researchers to positive results with Social Media data? However, some researchers are pessimist about Social Media analysis. According to (Ruths and Pfeffer, 2014; Weller, 2015), Social Media is noisy, and the data derived from it are of mixed quality: for every relevant post there may be millions that should be ignored. Learning with Social Media data sometimes requires robust statistical models. Nevertheless, researchers continue to investigate how best to make use of Social Media data. First studies show positive findings.

Social Media users not only react to and talk about events in real time, but also talk about and react to events that will happen in the future. This fact fuels the interesting possibility that Social Media data might be useful for forecasting events: making predictions about future events. Not only have researchers begun to investigate this line of questioning, earlier review articles on Social Media forecasting showcase early positive examples of predictive success (Kalampokis et al., 2013; O'Leary, 2015; Schoen et al., 2013). A lot of studies show that Social Media could be used to predict the future. At the same time, some works have been controversial (Schoen et al., 2013). It's clear that this domain of research is in its infancy, methodologies are different, common best practices are difficult to determine, and true replication of studies is near-impossible due to data sharing concerns (@ Weller, 2015). The use of data from Social Media for modelling real-world events and behavior has seen a growing interest since his first appearance in academic world around 2008. This increasing popularity is proportional to the leaps ahead made in computational social science. In the past, many sociological theories were hard to prove for the difficulties encountered in gathering indispensable data. Today, Social Media can record so many sides of human relationships on the web from millions of people all around the world. On the other hand, Social Media data cannot always provide a complete picture of what researchers might hope to see. The use of Social Media varies depending on age, culture, social background, gender and ethnicity. However, positive findings and the interest in fundamental dynamics of Social Media platforms explain the exponential growth in popularity of this field of research.

Social Media data has a huge potential but understanding if its application can be useful is not a trivial task.

Forecasting models (data- or theory-driven) are important in many fields but Social Media data challenges researchers to find new ways to apply them. In natural sciences, aggregating techniques of data coming from network of sensors are important, but Social Media data challenges researchers to find new ways to increase their forecasting power. Researchers should first identify the methods through which Social Media challenges may be addressed to be able to make valid and reliable predictions. Among these difficulties, there are: noisy data, possible biases, a rapidly shifting Social Media landscape that prevents generalization and a need for domain-specific theory that brings all together.

Furthermore it is important to chose the best text source for Social Media analysis, among the many available. Previous studies found that researchers focused mainly on Twitter data (Giacomo, 2017). While Facebook is trying to compete, and Snapchat offers a unique perspective on the theme, Twitter remains the best indicator of the wider pulse of the world and what is happening in it. According to Hamad (Ahmed, 2017), there are at least six reasons that explain the importance of Twitter for Social Media analysis: 1. Twitter is a popular platform in terms of the media attention it receives, and it therefore attracts more research due to its cultural status; 2. Twitter makes it easier to find and follow conversations (i.e., by both its search feature and by tweets appearing in Google search results); 3. Twitter has hashtag norms which make it easier gathering, sorting, and expanding searches when collecting data; 4. Twitter data is easy to retrieve as major incidents, news stories and events on Twitter are tending to be centered around a hashtag; 5. The Twitter API is more open and accessible compared to other Social Media platforms, which makes it more favorable to developers creating tools to access data. This consequently increases the availability of tools to researchers; 6. Many researchers themselves are using Twitter and because of their favorable personal experiences, they feel more comfortable with researching a familiar platform. It is probable that a combination of the response from 1 to 6 led to more research on Twitter. However, this raises another distinct but closely related question: when research is focused so heavily on Twitter, what (if any) are the implications of this on methods? As for the methods that are currently used in analysing Twitter data i.e., sentiment analysis, time series analysis (examining peaks in tweets), network analysis etc., can these be applied to other platforms or are different tools, methods and techniques required?

Below has to be considered whether these methods would work for other Social Media platforms (Ahmed, 2017):

1. Sentiment analysis works well with Twitter data, as tweets are consistent in length would sentiment analysis work well with, for example Facebook data where posts may be longer?
2. Time series analysis is normally used when examining tweets overtime to see when a peak of tweets may occur, would examining time stamps in Facebook posts, or Instagram posts, for example, produce the same results? Or is this only a viable method because of the real-time nature of Twitter data?
3. Network analysis is used to visualize the connections between people and to better understand the structure of the conversation. Would this work as well on other platforms whereby users may not be connected to each other i.e., public Facebook pages?
4. Machine learning methods may work well with Twitter data due to the length of tweets but would these work for longer posts and for platforms that are not text based, i.e., Instagram?

Maybe at least some of these methods can be applied to other platforms, however they may not be the best methods, and may require the formulation of new methods and tools. In conclusion, Twitter is the best for Social Media analysis for now. Despite its smaller user base compared with Facebook, its responsiveness and openness to researchers' tool make possible gathering useful data.

Since the usage of social media has a wide impact on a great number of disciplines, here is exposed the main literature in the most technical related fields that are strongly related to social media analysis: economics and marketing.

6.4.1 Economics

This domain has raised the great interest of researchers. The first studies focused especially on market fluctuation and on aggregated measure, such as Dow Jones Industrial Average (DJIA). Most recent researches have gone further predicting single stock price and yield.

Great interest in Social Media analysis for economics has been on Stock market analysis. Stock price forecasting is an important and thriving topic in financial engineering and is considered a very difficult task, even outside Social Media. Many articles in this context present models based on sentiment analysis to make forecasts (Xu and Keelj, 2014; Kordonis et al., 2016; Cakra and Trisedya, 2015,?; Wang and Wang, 2016; Shen et al., 2016; Brown, 2012; Rao and Srivastava, 2012), although some researchers realised more detailed models: Crone et al. (Crone and Koeppel, 2014) implemented neural networks and incorporated non-SM sources, and Shen et al. (Shen et al., 2016) developed a model that studies the connection between consumers' emotion and commodity prices.

The simplest task for stock market forecasting is predicting whether the following day will see rise or fall in stock prices. Comparison between researches is complicated by the fact that stock market volatility, and so the difficulty of prediction, may vary over time periods. High accuracy on this task was reported by Bollen et al. (Bollen et al., 2011), using sentiment analysis to achieve an accuracy of 87,6%. They investigated whether measurements of collective mood states derived from large-scale Twitter feeds are correlated to the value of the Dow Jones Industrial Average (DJIA) over time. They analysed the text content of daily Twitter feeds by two mood tracking tools, namely OpinionFinder, that measures positive vs negative mood, and Google-Profile of Mood States (GPOMS) that measures mood in terms of 6 dimensions (Calm, Alert, Sure, Vital, Kind, and Happy). They find that measures of "calm" on Twitter along with DJIA numbers from the previous three days provide the best up/down predictions. Further adding the emotion "happy" reduces rise/fall accuracy to 80% but does reduce error in terms of forecasting absolute DJIA values. Importantly, they find that positive/negative sentiment analysis through the popular OpinionFinder's tool leads to no improvement over just using previous DJIA values. In conclusion, researchers obtained good results forecasting up/down movements in the stock market.

Furthermore the topic of Sales and revenues is of great interest for people working on economics. For example, boosting movie ticket sales is an important task for producers and publishers, and this has been studied specifically on Social Media platforms like Twitter. Asur et al. (Asur and Huberman, 2010) showed how social media content can be used to predict movie success. In particular, they used the chatter from Twitter.com to forecast box-office revenues for movies. Specifically, using the rate of chatter from almost 3 million tweets, they constructed a linear regression model for predicting box-office revenues of movies in advance of their release. Then, they showed that the results outperformed in accuracy those of the Hollywood Stock Exchange and that there is a strong correlation between the amount of attention a given topic has (in this case a forthcoming movie) and its ranking in the future. They also analysed the sentiments present in tweets and demonstrated their efficacy at improving predictions after a movie has released. Cheng et al. (Cheng et al., 2013) obtained mixed results developing a model for predicting TV audience rating. They accumulated the broadcasted TV programs' word-of-mouse on Facebook and apply the Back-propagation Network to predict the latest program audience rating. They also presented the audience rating trend analysis on demo system which is used to describe the relation between predictive audience rating and Nielsen TV rating. Kim et al. (Kim et al., 2014) investigated the relationship between music listening behavior in Twitter and the Billboard rankings. They found that the play-counts extracted from tweets have strong relationships with the Billboard rank, whereas, interestingly, the artist popularity extracted from tweets has a weak correlation with future chart rankings. In addition, the number of weeks on chart information alone was insufficient to predict rank alone. With the features extracted from tweets, They built three regression models to predict the ranking. Among the proposed models, SVR (Support Vector Machine) showed the highest squared correlation coefficient (0.75). Although the combined model with the number of weeks on chart performed the best in rank prediction, the music listening behavior available in Twitter can generate an outstanding predictive model. They also built a hit prediction classifier with the features acquired in tweets and the number of weeks on chart. They classified the hit and non-hit songs in the Billboard Hot 100 and obtained a value of 83.9% accuracy, 83% precision, and 85.3% recall for classifying a hit song over the whole data set. The proposed feature showed a high performance both for rank prediction and hit classification. The previous week's twitter features and the number of weeks on chart are effective for predicting the Billboard rank of a song. Ahn et al. (Ahn and Spangler, 2014) focused on periodic forecasting problems of product sales based on social media analysis and time-series analysis. In particular, they presented a predictive model of monthly automobile sales using sentiment and topical keyword frequencies related to the target brand over time on social media. Their predictive model illustrates

how different time scale-based predictors derived from sentiment and topical keyword frequencies can improve the prediction of the future sales. Tuarob et al. (Tuarob and Tucker, 2013) proposed a Knowledge Discovery in Databases (KDD) model for predicting product market adoption and longevity using large scale, social media data. In particular, the authors analysed the sentiment in tweets and use the results to predict product sales. The authors presented a mathematical model that can quantify the correlations between social media sentiment and product market adoption in an effort to compute the ability to stay in the market of individual products. The proposed technique involves computing the Subjectivity, Polarity, and Favorability of the product. Finally, the authors utilised Information Retrieval techniques to mine users' opinions about strong, weak, and controversial features of a given product model. The authors evaluated their approaches using the real-world smartphone data, which are obtained from www.statista.com and www.gsmarena.com. The findings show that tweets can be used to predict product sales for up to at least 3 months in advance for well-known products such as Apple iPhone 4, Samsung Galaxy S 4G, and Samsung Galaxy S II, thus the predictive ability varies across products.

6.4.2 Marketing

Scholars had a great focus in the last years on using Social Media Information for marketing. Chen et al. (Chen et al., 2015) conducted a survey study and a field study to explore the feasibility of using predicted personality traits derived from social media text for the purpose of ad targeting. In the survey study, they measured people's personalities and their responses to an advertisement tweet. They found that people with high openness and low neuroticism responded more favorably to a targeted advertisement, thus demonstrating the effects of the personality traits themselves. In the field study, they sent the advertisement tweets to real-world Twitter users, and found the same effects on users' responses using personality traits derived from users' tweet text. They demonstrate that aiming advertisements at users with particular personality traits improves click and follow rates by 66% and 87% respectively, representing a large increase in value for companies. These results suggest that the derived personality traits had the same effects as the personality traits measured by traditional personality questionnaires and can indeed improve ad targeting in real-world settings. Li et al. (Li et al., 2016) present a solution to the problem of predicting project success in a crowd-funding environment combined with innovative introduction of survival analysis based approaches. They used comprehensive data of 18 thousand Kick-starter (a popular crowd-funding platform) projects and 116 thousand corresponding tweets collected from Twitter. While the day of success is considered to be the time to reach an event, the failed projects are considered to be censored since the day of success is not known. They performed rigorous analysis of the Kick-starter crowd-funding domain to reveal unique insights about factors that impact the success of projects. Their experimental results show that incorporation of failed projects (censored information) can significantly help in building a robust prediction model. Additionally, they also created several Twitter-based features to study the impact of social network on the crowd-funding domain. Their study shows that these social network-based features can help in improving the prediction performance. They found that the temporal features obtained at the beginning stage (first 3 days) of each project will significantly improve the prediction performance. Even when just using Social Media information from the first three days of the project, they achieve an AUC of 0.90, reflecting very high classification performance.

Part III

Methods and Results

L'aprrocio metodologico generico…

This part describes the methods applied for the analysis of technical documents. The methods are ensamble of Natural Language Processing (NLP) and Text Mining *techniques* described in 5, re-designed depending on the analyzed document and the analysis goal. Not all the *techniques* have been applied to all the documents: table tot summarise the relations between the documents under analysis (introduced in section 6) and the NLP techniques.

Table documents vs tools

Each chapter starts with a brief description of the field of applicaton of the method and with the framing of the problem to be solved. Then the methodlogy to solve the prolbem is described. Each chapter closes with the results.

Chapter 7

Patents

Patents contain a large quantity of information which is usually neglected. This information is hidden beneath technical and juridical jargon and therefore so many potential readers cannot take advantage of it. State of the art natural language processing tools and in particular named entity recognition tools, could be used to detect valuable concepts in patent documents. A deeper description of what patents are and how these documents are used to mine technical knowledge can be found in section 6.1

In this section we present three methodologies capable of automatically detecting and extracting threee of the multiple entities hidden in patents: the users of the invention, advantages and drawbacks of the invention and trademarks contained in patents. The results of the methodologies are described, togheter with example of applications of the extracted entities for intelligence tasks.

7.1 Users

Patents are documents that must provide a detailed public disclosure of an invention (Idris, 2008). An *invention* is a new and useful process, machine, manufacture, or composition of matter, or any new and useful improvement thereof [1].

The notion of usefulness implies that the invention must have some value and not necessarily for a human entity. In fact, patents usually describe processes, machines or composition of matter which are useful for another process, machine or composition of matter.

Therefore, we distinguish between stakeholders and users, considering the definitions given by the authors in (Bonaccorsi and D'amico, 2017).

Definition 1: **Stakeholder** : *Stakeholders are entities on which the invention has or will have a positive or negative effect in order to show usefulness.*

This definition covers all possible entities that engage an active or passive relation with the invention.Given the logical condition of usefulness of patents, all patents must have stakeholder information. If a patent has not got any stakeholder information in it the patent application should be rejected.

Definition 2: **User**: *Users are animated or previously animated entities (human or animal, alive or dead), on which the invention has a positive or negative effect at an unspecified moment.*

Given definition 2, it is clear that every user is a stakeholder while non-users stakeholders include artifacts, machines, manufacturing or operational processes.

Corollary 1: **Multiple roles**: *Identities may have multiple roles as users.*

[1] http://www.uspto.gov

Our idea of users describes roles, not identities. Animated entities have an identity, as it happens for a specific person. A person has many roles as a user. For example, a *working mother* starts her day taking on the role of a *mom*, in which she is expected to feed her children and get them ready for school. At the office she shifts to the role of *project manager*, so she oversees projects in a timely and professional manner. *Working mother*, *mom* or *project manager* can be considered user roles attributed to the same person, or identity. From this definition it is clear that users are close to what social sciences define as "social roles".

Afterward, we can outline knowledge fields using the concept of user, with a twofold aim: help the reader to understand how the concept of users is interpreted in different knowledge fields; explain the background of the methodology we adopted.

Social sciences: social roles as users

In social sciences *social roles* are comparable with our definition of user. As defined in the psychological field (Dog, 2015), *"social roles refer to the expectations, responsibilities and behaviors we adopt in certain situations."*. The example of the working mother shown before, is the case of social roles.

The field of social sciences is the only one in which an attempt of automatic extraction of users has been done. In (Beller et al., 2014b) the authors extracted social roles from Twitter using heuristic methods. The authors looked for all the words preceded by constructions like "I'm a" and similar variations. This search resulted in 63.858 unique roles identified, 44.260 of which appeared only once. The result of the extraction process is noisy and only a low percentage of the extracted words are social roles. Despite of this noisy extraction, some entities are consistent with our definition of user, e.g. *doctor, teacher*, *mother* or *christian*.

Another work (Beller et al., 2014a) tries to identify social roles on Twitter exploiting a set of assumptions. The authors take into account roles, each one with a set of related verbs: if someone uses verbs from a set, that person may cover that particular social role. To sanitize the collection of positively identified users, the authors crowd-sourced a manual verification procedure, using the Mechanical Turk platform (Kittur et al., 2008). Also here some interesting extractions are performed, obtaining users like *artist, athlete, blogger, cheerleader, christian, DJ*, or *filmmaker*. These two works differ from the present study for what concerns the analyzed texts and the methods to extract the entities. Nevertheless, the extracted set of entities is consistent with our definition of user.

Human Resources Management: workers as users

In organizations, Human Resources Management is the function designed to maximize employees performance (Johnson, 2009). Employees are key actors and they can be considered users according to our definition.

Human Resources Management has tried to classify employees, especially in sub-fields like insurance, social security or work psychology. Usually, we refer to those as lists of jobs. Classifications were made with the goal of grouping similar jobs for educational requirements, job outlooks, salary ranges or work environments to facilitate social analysis and the placement of new workers. Such lists are relevant because, even if they represent just one subset of all the possible users, they contain valid information. Many institutions developed lists of jobs (lis, 1967).

Medicine: patients as users

Another field of interest is medicine, since patients can be considered users. Also in this case there are many lists of patients, illnesses and diseases (of Health and Services, 2018), which are valuable in terms of information contained.

Design and Marketing: between users and customers

In the field of *Design* the concept of user plays a central role and it overlaps with our definition of user. Many tools and theories like "User Centered Design" are based on the concept of user (ISO, 1999). As stated by the authors in (K., 2008), the quality of the design process is proportional to the user needs' satisfaction. It implies that a designer has to understand the user needs; as a consequence he has to discover whom are potential users.

7.1.1 Method

In this section we show the approach used to extract the users of the invention described in a patent. The proposed process is shown in figure 7.1 and its phases are:

1. *Generation of an input list of users*: search all possible sources with the aim of creating an input list of users with the largest possible coverage (section **??**);
2. *Patent set selection*: select the set of documents from which extract the users (section **??**);
3. *Patent text pre-processing*: application of natural language processing tools on the documents with the aim of preparing them for the automatic user extraction;
4. *Automatic patent set annotation 1*: projection of the input list of users on the text to generate the Automatically Annotated Patent Set 1;
5. *Relevant sentences extraction*: selection of sentences containing at least one user to generate an informative training set;
6. *Automatic patent set annotation 2*: generation of a statistical model by a machine learning algorithm based on the training set sentences and automatically tagging the patent set to generate the Automatically Annotated Patent Set 2;
7. *Difference computation*: generation of the new list of users by computing the difference between the lists of users found in the automatically annotated patent set 1 and 2;
8. *Manual review*: manual selection of the entities that, in the new list of users, are effectively users. This new list will enrich the original list of users. This phase is described in section **??**.

Before the description of each phases, in section **??** the concept of *user of the invetion* is explained by giving a definition of users and presenting the way that this concept is exploited in different knowledge fields.

List of users generation

To generate the input list of users, we used two different approaches: a bottom-up approach and a top-down approach. The bottom-up approach is based on the merge of lists from heterogeneous sources. In the present work we used the following lists of entities:

- *Lists of jobs* :(lis, 1967), 11.142 entities
- *Lists of sports and hobbies*[2]: 9.660 entities;

List of animals [3]: 600 entities;

- *Lists of patients* [4]: 14.609 users;
- *List of generic words*: manually generated. It contains users with a higher level of abstraction (such as *person* or *human being*), 56 items.

[2]http://www.notsoboringlife.com/list-of-hobbies/,http://www.notsoboringlife.com/list-of-hobbies/
[3]http://a-z-animals.com/animals/
[4]http://www.medicinenet.com/diseases/_and/_conditions/alpha/_a.htm, http://www.cdc.gov/DiseasesConditions/az/a.html

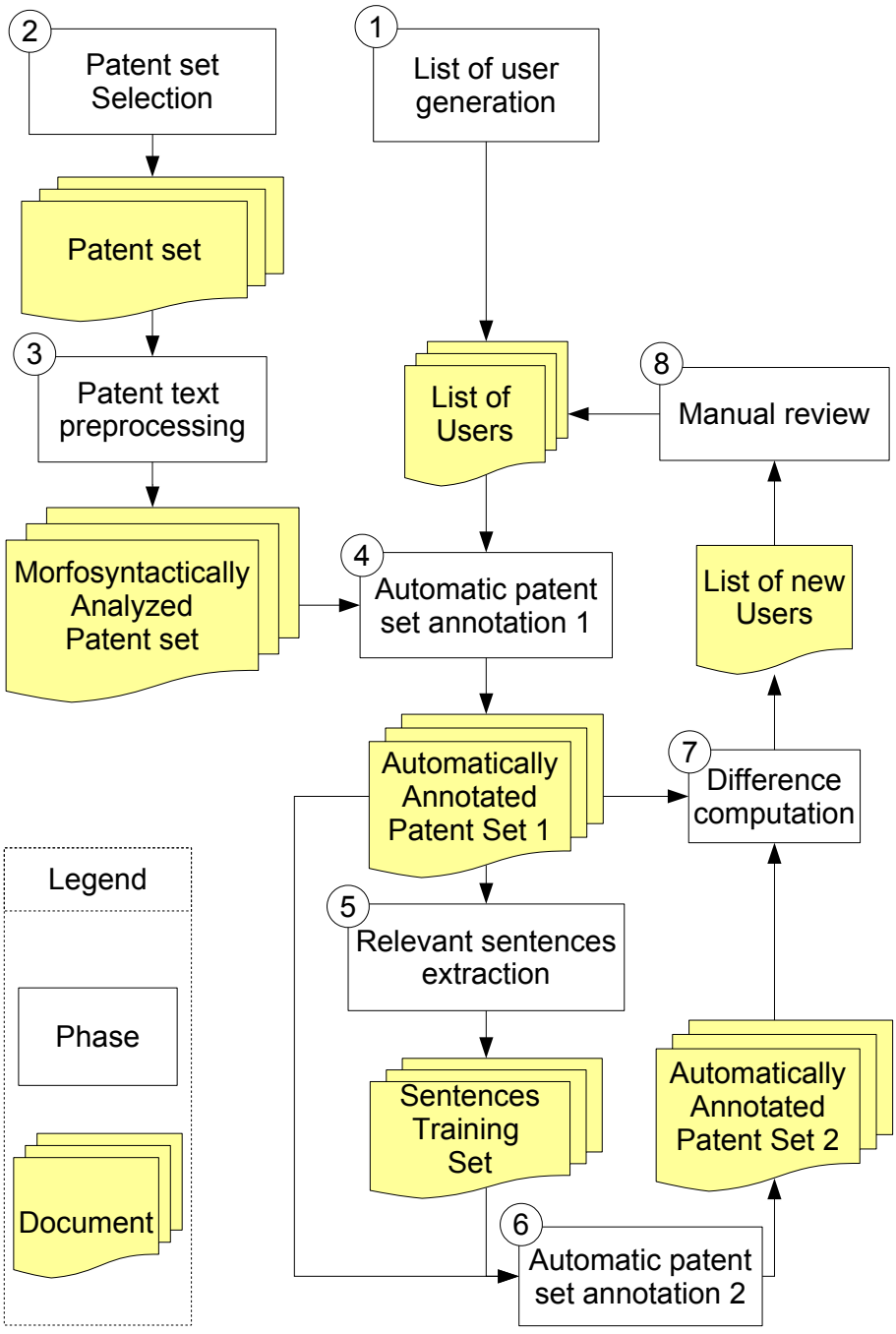

Figure 7.1: Process flow diagram of the proposed automatic user extraction system from patents. The diagram contains the representation of the documents and the operations performed on them. The process takes in input a patent set and a list of users and produces a list of new users as output.

Bottom-up approach produced a list of 35.767 entries.

Afterwards, a top-down approach was applied. Starting from the list generated with the bottom-up approach, we looked for alternative methods to indicate a user, finding defined word patterns. The most relevant are:

- Patterns like "hobby_term + practitioner" for the hobbies;
- Patterns like "person who has + disease_term" or "suffering from + disease_term" for the diseases;
- Patterns like "practitioner of + sport_term" for sports.

Top-down approach generated a total of 41.090 entries.

The whole process generated a total of 76.857 users and gave us a reasonable number of terms to be used in the next step of the process.

Obviously our lists have a limited coverage and, therefore, they do not contain all variations of a certain user. For instance, the lists miss some users belonging to the classes mentioned above (e.g. new jobs emerged in the last years) and all the alternative ways for referring to a user we do not spotted in the top-down approach. For example our lists miss jobs like *data analyst*, *lap dancer*, *undertaker*, *mortician* and *thief* or patients with emerging diseases like *work-alcoholic* and *web-addicted*. In addition, our lists miss a class of users related to religious groups, containing users like *christians* or *jewish*. Such terms have intentionally **not** been introduced in the input list because we considered these terms as candidates to be extracted by the process in our case study .

Patent set selection

Our choice of patent sets aimed at challenging our system to find new users missing in the input list. To reproduce a patent set selection, we took into consideration the International Patent Classification (IPC) (Organization, 1971). IPC is a hierarchical system of patent classes representing different areas of technology. Then, we wondered which classes could contain new users according to our seed list. Furthermore, IPC class A, which is the first level in IPC differentiation, is based on human necessities. For this reason, we assumed that in this class we would have found likely users from patents texts.

Patent text analysis

Our Entity Extraction system is composed by a set of sequential phases. The first three phases are related to the linguistic annotation: sentence splitting and tokenization, part of speech tagging and lemmatization. Then, the patent set is analyzed by the entity extractor, specialized for users extraction. A more detailed description of each phase is:

- Sentence splitting and Tokenization: These processes split the text into sentences and then segment each sentence in orthographic units called tokens. In our system, sentence splitting plays a key role since thanks to a given word, it is possible to find sentences where the word is used. Finding correct boundaries for a specific word allows to dramatically reduce the space to retrieve its surrounding contexts.
- POS tagging and Lemmatization: The Part-Of-Speech tagging (or POS tagging) is the process of assigning unambiguous grammatical categories to words in context. It plays a key role in NLP and in many language technology systems. For the present application we used the most recent version of the Felice-POS-tagger described in (Dell'Orletta, 2009). Once the computation of the POS-tagged text is completed, the text is lemmatized according to the result of this analysis.
- Semi-automatic Users Annotation: The Users Extraction tool is based on supervised methods. Such methods require an entity annotated corpus in order to extract new entities from unseen documents. A semi-automatic method has been used to generate an annotated corpus of users to avoid manual annotation of a patent set. The method is a projection of the list of users on the patent set defined in section @ref{patsel}. The list of users described in section @ref{sourc} is cleaned to avoid linguistic ambiguities when projecting these entities on the corpus. For example, the term *"guide"* has two different meanings when used as a verb or as a noun. Furthermore, as a noun it could indicate a

component of a system (guide for mechanical parts) or a person (someone employed to conduct others) and therefore a user. Avoiding ambiguities is a crucial aspect to produce an informative training set, so ambiguous words were pruned.

The entity annotation schema for a single token is defined using a widely accepted BIO annotation scheme Ramshaw and Marcus (1999):

- **B-USE**: the token is the beginning of an entity representing an User;
- **I-USE**: the token is the continuation of a sequence of tokens representing an User;
- **O**: for all the other cases.

User Entity Extraction

The Users extraction problem is tackled by the implementation of a supervised classifier that is trained on an annotated patent set. Thus, the patent set is linguistically-annotated, using the steps described above and entity-annotated, exploiting the semiautomatic annotation process executed in the previous steps.

Given a set of features the classifier trains a statistical model using the feature statistics extracted from the corpus. For each new document the trained model assigns to each word the probability of belonging to one of the classes previously defined (B-USE, I-USE, O).

In our experiments the classifier has been trained using two different learning algorithms: Support Vector Machines (SVM) using the LIBSVM library (Hearst et al., 1998a) configured to use a linear kernel and Multi Layer Perceptron (MLP) implemented using the Keras library (Chollet, 2015). It has been proven that LSTM methods are well suited for similar NER task. Anyway, we chose SVM and MLP method to study how two wheel established state of the art classifiers perform on the specific task of user extraction from patents and to evaluate their performance in terms of precision and computational effort. We also think that the popularity of these methods increment the reproducibility of the work.

The classifier uses different kind of features extracted from the text:

- *linguistic features*, i.e. lemma, Part-Of-Speech, prefix and suffix of the analyzed token;
- *contextual features*, the linguistic characteristics of the context words of the analyzed token; in addition the entity category of the previous token is considered;
- *compositional features*, combinations of contextual features and linguistic features. i.e. Part-Of-Speech of the previous word and the lemma of the current word. These extra features allow to infer statistics on the interaction of the combined features that can not be captured by a linear SVM model.
- *word2vec features*: vector representations of words computed by the *word2vec* (Mikolov et al., 2013) tool.

Word2vec is a NLP tool able to produce word representations exploiting big corpora. The main property of the vectors produced by *word2vec* is that words sharing similar contexts have similar vector representations. By using word vectors instead of the corresponding words we were able overcome the problem of the limited lexical knowledge in the training phase. Using these features and excluding all the others (delexicalized model) we expected that the resulting user extraction system had a lower precision and an higher recall in the classification phase. We presumed to find new users not contained in the input seed list.

Manual Review of the new list of users

It is still possible that the classification process creates false positive results (words labeled as users that do not match the definition in section @ref{theuse}). Thus, it is necessary to make a manual review of the extracted entities with the aim of evaluating the output.

7.1.2 Results

The following section describes the performances of the automatic users extraction process on two different patent sets. To test the system four experiments were conducted}. Finally the performances and the outcomes of the system are shown and discussed.

Following the guidelines for the patent set selection described in section **??**, we examined two patent sets belonging to the IPC class A:

- **A47G33**. The IPC definition of the subclass is *"religious or ritual equipment in dwelling or for general"*.
- **A61G1-A61G13**. The IPC definition of the subclass A61G1 is *"Stretchers"* while the definition of the subclass A61G13 is *"Operating tables; Auxiliary appliances therefor"*.

We extracted from the private Errequadro s.r.l. [5] database a random sample of 2.000 patents from each IPC class. For each patent set we applied the semiautomatic set annotation process by projecting the input list of users on the morphosyntactically analyzed patent set. After this process, each semi-automatically annotated patent set was split in two parts: the first was used as training set for the user extractor, and the second one was used as test set.

To build an informative training set, from the semi-automatically patent set we selected a subset of sentences containing at least one user. The size of the training set in both cases is approximately composed by 600.000 tokens. For each patent set table 7.1 shows the number of sentences of the training set, the number of sentences of the test set, and the number of distinct users in the training set (re-projected by the semi-automatic annotation process).

ref —> patent set-details

Table 7.1: Statistics related to the patent set groups analyzed in the case study

patent set group	*#Sentences - training*	*#Sentences - test*	*#Distinct users projected on training*
A47G33	13.364	214.029	126
A61G1-A61G13	15.108	2.520.350	121

We chose two orders of magnitude for the sentences test-set to test the efficiency of multiple configurations of the system.

To test the performances of the implemented user extractor, we devised four different configurations. Each configuration uses a specific learning algorithm and a set of features to build the statistical model. The main purpose of this procedure is to find the configurations that better perform in the user extraction task. In addition, the different behaviour of the system in the classification phase is studied. In table 7.2 are reported the detailed configurations used in our experiments.

Table 7.2: Context windows of the extracted features considering 0 as the current analyzed token.

Feature group	*Context Window*
Lemma unigrams	\([-2, -1, 0, 1]\)
Lemma bigrams	\([(-1 ,0), (0, 1)]\)
Word bigrams	\([(-1 ,0), (-2, -1), (0, 1), (1, 2)]\)
Word trigrams	\([(1, 0, 1) (-2, 1, 0)]\)
Pos unigrams	\([-2, -1, 0, 1]\)

[5] http://www.errequadrosrl.com/

Feature group	*Context Window*
Pos bigrams	\(([(-2, -1) (-1, 0), (0,1)])\)
Compositional feature #1	\((POS_{-1}, Lemma_{0})\)
Compositional feature #2	\((Lemma_{-1}, Lemma_{0})\)
Compositional feature #3	\((Lemma_{0}, Lemma_{1})\)
Compositional feature #4	\((POS_{0}, Lemma_{1})\)
Compositional feature #5	\((NER_{-1}, Lemma_{0})\)
Word2vec	$-2, -1, 0, 1, 2$

By using the first and the second configuration we expected to have a higher precision in the classification phase, since explicit lexical information is used in the training phase. For the same reason we expected to have low recall in classification phase. On the other hand, the third and fourth configurations are delexicalized: lexical information is provided by word vectors computed by word2vec_. In these two configurations we expected to have an higher recall and a lower precision, due to the characteristics of the computed vectors explained before. To limit errors when using the *word2vec* features, some linguistically motivated filtering rules were introduced. Specifically, sequences of tokens classified as users were constrained from the following categories: verbs, adjectives not preceded by articles, articles and adverbs.

To evaluate the whole user extraction process in each experiment, we defined some evaluation measures. Each measure was introduced to evaluate the characteristics of the extraction system concerning the configuration applied.

These measures are:

- Training time: time needed to create the statistical model using the training set;
- Test time: time needed to re-annotate the semi-automatically annotated patent set;
- Number of extracted users: number of unique entities classified as user in the automatically annotated patent set;
- Number of known users: number of distinct extracted users in the automatically annotated patent set and belonging to the list of user in input;
- Number of new users: number of distinct entities classified as user in the automatically annotated patent set and not belonging to the input list of users;
- Number of new correct users: number of distinct entities considered as user and as correct after a manual review;
- Precision: ratio between the number of new distinct correct users and the total number of new distinct users;
- Gain: ratio between the number of new distinct correct users and the number of re-projected distinct users on the training set.

Table 7.3 reports the values of the defined metrics across all the experiments run on the two patent sets.

Table 7.3: Comparison of the values of the defined metrics across all the experiments. The patent set annotation in the experiment (6) was not performed due to the computational costs. All the experiments were run on a machine provided with 10 AMD Opteron(tm) 6376 processors.

Experiment	Training time	Test Time	Extracted	Known	New	New correct	New wrong	Prec. (%)	Gain (%)
1 (SVM)	83m	321m	161	93	68	47	21	69.11	37.30
2 (MLP)	1911m	9091m	196	55	141	27	114	19.15	21.42
3 (MLP-W2V)	165m	246m	162	35	127	45	82	35.43	35.71

Experiment	Training time	Test Time	Extracted	Known	New	New correct	New wrong	Prec. (%)	Gain (%)
4 (SVM-W2V)	1265m	4310m	121	29	92	45	47	48.91	35.71
5 (SVM)	148m	3443m	302	120	182	88	108	48.35	72.72
6 (MLP)	1818m	—	—	—	—	—	—	—	—
7 (MLP-W2V)	333m	3530m	305	38	267	44	230	16.48	36.36
8 (SVM-W2V)	1268m	47020m	313	49	264	74	197	28.03	61.15

For what concerns training and test time of the automatic patent set annotation, it's clear that the configuration based on the SVM learning algorithm without the *word2vec* features performs better in both the experiments (1, 5). When the features based on *word2vec* are introduced, the configuration based on the MLP learning algorithm is the fastest both in training and test time (3, 6): it is due to the fact that keras implementation of this algorithm exploits all the available CPU cores of the system. On the other side, the MLP algorithm does not scale properly with a higher number of features, as seen in training and annotation time in the experiment (2). In addition, we could not perform the patent set annotation in the experiment (6), since it would have required more than 60 machine days to complete the process. When *word2vec* features are introduced, the patent set annotation based on the SVM algorithm is 10 times slower than the MLP algorithm.

For what concerns the precision in the automatic patent set annotation, the SVM configuration without *word2vec* features is clearly the more reliable: the precision values are from 1.5 to 2 times higher in the experiments (1, 5) in contrast to the other experiments. The higher precision is justified by the fact that the configurations based on *word2vec* features lack explicit lexical information: words with very similar contexts are represented by similar *word2vec* vectors, probably leading to errors in the classification phase. On the other hand, the use of *word2vec* vectors aims at extracting entities that would not be extracted by considering explicit lexical information only.

Finally, for what concerns information gain, the same amount of new information (21-37%) is extracted in the experiments on the A47G33 patent set. The gain values drastically change in the experiments on the A61G1-A61G13 patent set: in the experiments (5, 8) a gain between 61% and 72% is obtained: it is due to the size of this patent set in comparison to the A47G33 one. In the experiment (7), despite the introduction of *word2vec* features, a gain of 36% is obtained. This fact, in conjunction with the non-feasibility of the experimental configuration 6, shows how MLP systems lack in efficacy and efficiency (in entity extraction in patent domain) when the test-set has an order of magnitude of millions of sentences. We think that this result is relevant, based on our experience with practical applications.

Furthremore, a way to maximize the overall informative gain is to merge the results of all manually reviewed user extractions obtained by executing the patent set annotation process with all possible configurations.

The overall informative gain of the merging process is related to intersections that occur among the results obtained by the patent set annotation process in each configuration: the less the intersections, the more the overall informative gain obtained. In table 7.4 is shown the overall gain obtained by merging results of the manually reviewed extractions in each patent set.

Table 7.4: Gain obtained by merging correct entities extracted from each patent set annotation.

Configuration	A47G33 - Gain (%)	A61G1+A61G11 - Gain (%)
SVM	37.30	72.72
MLP	21.42	—
MLP-W2V	35.71	36.36

Configuration	A47G33 - Gain (%)	A61G1+A61G11 - Gain (%)
SVM-W2V	35.71	61.15
SVM - MLP	52.38	—
SVM - MLP-W2V	69.84	126.44
SVM - SVM-W2V	73.01	103.30
MLP - MLP-W2V	55.55	—
MLP - SVM-W2V	57.14	—
MLP-W2V - SVM-W2V	59.52	76.30
SVM - SVM-W2V - MLP-W2V	90.47	140.49
SVM - MLP - MLP-W2V	82.53	—
SVM - MLP - SVM-W2V	85.71	—
MLP - SVM-W2V - MLP-W2V	77.77	—
SVM - MLP - SVM-W2V - MLP-W2V	103.17	—

The table shows that the merging process of manually reviewed entities extracted from each patent set annotation run effectively contributes to increase the overall informative gain. For instance in the A47G33 patent set an overall gain of 103.17% is obtained, tripling the best result achieved by the extraction performed using the best single configuration. Good results are also achieved in the A47G33 patent set user extraction. In this case an overall gain of 140.49% is obtained, doubling the best result achieved by the extraction performed using the best single configuration.

The results shown in section 5 prove that if the goal of the extraction is to reach the maximal recall, an ensemble method (combining the output of multiple classifier) over-performs every single classifier method. Anyway, the ensemble approach has clear efficiency issues, because the time of analysis will be the sum of every single approach time (in hypotheses of non-parallelization). This leads to a trade off between the speed of the system and the quality of the results, and whoever would use the presented system can decide to gain benefit in one or in another direction.

Finally, tables 7.5 and 7.6 show an overview of extracted users randomly chosen from the A47G33 patent set (the only one in which were able to perform all experiments). Each table is divided in two blocks, representing the results of the extraction performed using a specific configuration. For each extracted user is shown the corresponding lemma (the root form), the frequency (how many times that user appears in the whole corpus) and the total number of patents containing the user. Users not contained in the starting user list, are highlighted in bold.

The table shows that the system was able to extract characteristic users of the patent set. The results are in fact not unexpected for the IPC class under analysis: this is an evidence of the correct performances of the proposed system. In other words, the results presented in the table show that it is possible to train a NER systems able to extract sparse and valuable information. Such users are the ones that an expert would manually extract but the NER system does it with an enormous saving in terms of time and efforts.

Other remarkable results are:

- many newly extracted entities have very low frequency in the patent set: it shows that the developed system is able to extract rare entities.
- table 7.6 shows that configurations using *word2vec* features are able to find new users with a higher frequency in the patent set: it was an expected result, since the *word2vec* configurations are not explicitly lexicalized and more able to generalize during extraction phase.
- The system is able to extract single words and multi-words.
- Taking into consideration the definition of user of an invention, the system extracts unusual and sometimes borderline users. Examples like *saint*, *angel*, *god* and *ghost* need discussion that is far beyond the purposes of the present paper. These results are a remarkable evidence of the human-like generalization ability of the described method.

Table 7.5: Extracted users from the A47G33 patent set using the SVM and DL configurations. New users extracted by the system are reported in bold.

Lemma	Frequency	# Patents	Lemma	Frequency	# Patents
female	801	109	child	402	102
child	426	108	cleregy member	128	5
guy	156	17	patient	113	11
patient	115	11	man	50	26
parent	70	31	young	48	32
man	51	26	**angel**	29	23
merchant	50	6	dog	20	7
soon	46	29	artisan	12	12
engineer	45	45	**male/female**	12	4
adult	39	23	hockey player	7	1
young	35	24	**professional**	7	7
society	32	21	tennis player	7	4
angel	29	23	football player	6	3
fund raiser	27	4	**ghost**	5	3
priest	22	4	children	5	5
cheerleader	15	4	manager	5	5
fund-raiser	11	4	**spider**	5	5
athlete	10	9	**vandal**	5	1
ghost	5	5	**athlete**	4	3
adulterant	3	3	mother	4	2
jew	3	3	soccer player	4	3
maid	3	1	squirrel	3	2
tourist	3	3	**maid**	3	1
indian	2	2	**god**	3	2
beginner	1	1	**mariner**	3	3
christians	1	1	**male-female**	2	2
datum entry operator	1	1	**manufacturer**	2	2
expert	1	1	**jew**	1	1
jewish	1	1	**merchandizers**	1	1
marinaro	1	1	**parishioner**	1	1

Table 7.6: Extracted users from the A47G33 patent set using the SVM-W2V and MLP-W2V configurations. New users extracted by the system are reported in bold.

Lemma	Frequency	# Patents	Lemma	Frequency	# Patents
child	152	68	clergy member	124	5
clergy member	124	5	**crowd**	36	3
man	50	26	basketball player	20	5
engineer	45	45	**him**	17	8
young	29	24	woman	16	8
choir	17	1	**saint**	14	2
infirm	13	8	**youth**	14	2
bride	9	4	**angel**	8	4
volunteer	8	6	**choir**	8	1
musician	6	6	musician	6	6
boy	3	1	**god**	5	1
children	3	3	children	3	3

girl	3	2	guy	3	3
creature	2	1	infant	3	3
deceased	2	1	priest	3	3
jewish	2	2	**bride**	2	2
person	2	2	**consumer**	2	2
mother	2	2	**everyone**	2	2
audience	1	1	**him/her**	2	2
boyfriend	1	1	**spectator**	2	2
derby member	1	1	farmer	2	1
gift giver	1	1	youngster	2	2
handicapped	1	1	**boyfriend**	1	1
jesus	1	1	**grandparent**	1	1
saint	1	1	**subject**	1	1
husband	1	1	clown	1	1
lady	1	1	husband	1	1
runner	1	1	runner	1	1
society	1	1	society	1	1
teenager	1	1	tennis player	1	1

The total number of users is 109. 28,2% (564 on 2.000) of patents in analysis contains at least one user. This result is an evidence of the fact that patents actually contain users information, and, considering the approach we followed, this percentage is an accurate lower approximation of the actual percentage of patents containing at least one user.

In figure 7.2 for each user on the x axes is shown the number of patents in which the user is contained. The distribution is skewed, with some occurrences showing large numbers and many others with just one or few occurrences. It is clear that there is a Pareto like distribution, with the first 20% of users covering 70% of total users in terms of occurrence. It means that some users are more likely to be cited in patents and many more users that rarely appear. Following this observations, we can divide users in three groups:

- *Group A*: users that appear in more than 100 patents (5% of the patent set). In our case these are *male*, *child* and *female*.
- *Group B*: users that appear in more than 20 patents (1% of the patent set). This group is composed by 13 different users. Some of these are *engineer*, *person*, *player*, *adult*, *angel* and _guy.
- *Group C*: users that appear in less then 20 patents. This group is composed by 93 different users. Some of these are *mother*, *athlete*, *priest*, *adulterant*, *golfer* and *hockey player*.

Further research means to study how these users differ from patent set to patent set. We expect to see similar distribution but with different content of users. Frequent and non-specific users comprise Group A: in other patent set we could see differences in terms of entities contained in this class but its content will stay non-specific. These results seem to be generic social roles indicating the gender or the age of a person. Group B is composed of mainly non-specific users and some specific users that change from patent set to patent set. This class helps to identify the core users of the patent set. Lastly, Group C contains non-frequent users that are both specific and non-specific, making it the most interesting of the three for the purposes of our work. In this group we find users that are market niches, so the patent that contains these users is of great interest for marketers and designers. These are both samples of the more generic users (for example a *mother* is a *female* and a *hockey player* is a *player*) or specific users of the patent-set (like *priest*, *fund-raiser*, *doll*, *spouse* and *clergy member*.

7.2 Advantages and Drawbacks

An effective development of new products or the redesign of an existing one require the analysis of its positive and negative properties. Due to that, advantages and drawbacks of products are extremely valuable

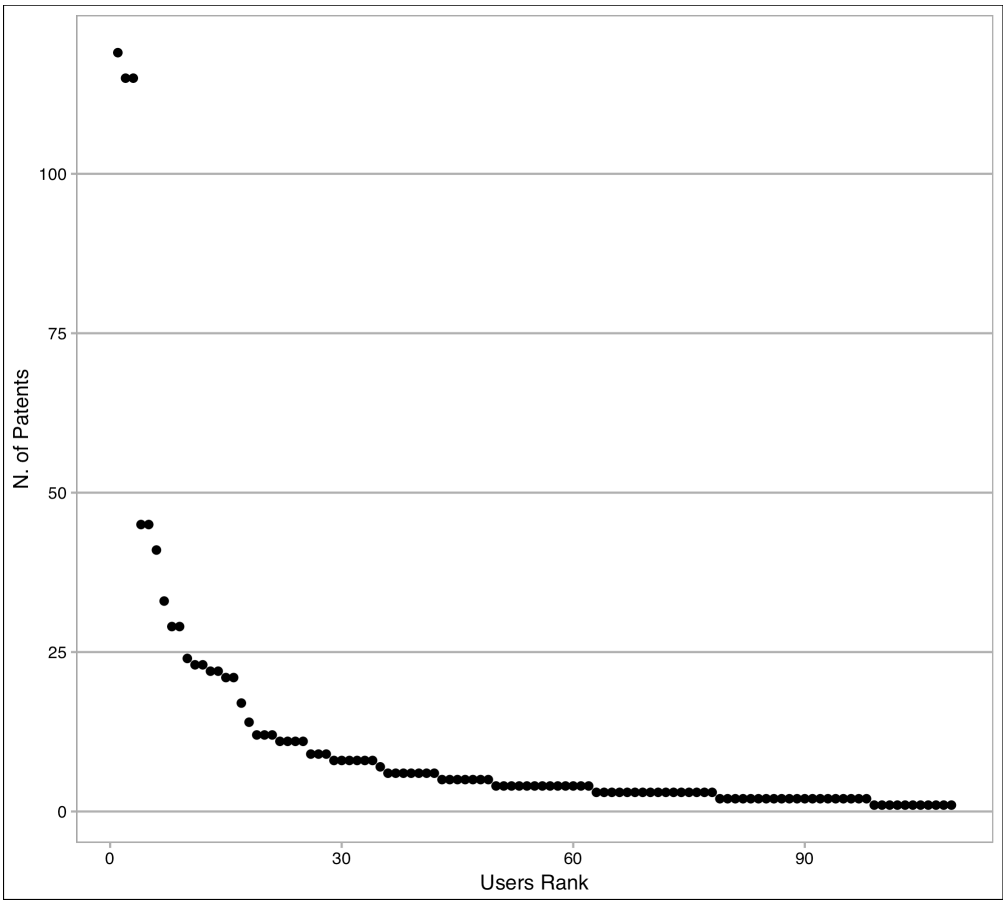

Figure 7.2: Process flow diagram of the proposed automatic user extraction system from patents. The diagram contains the representation of the documents and the operations performed on them. The process takes in input a patent set and a list of users and produces a list of new users as output.

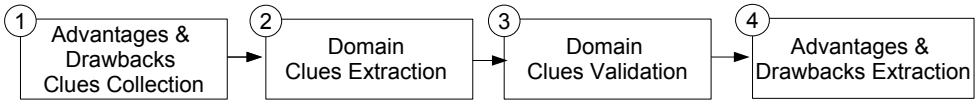

Figure 7.3: Main overview of the advantages and drawbacks extraction process from patents.

information for companies. Unfortunately, this information is not easy to obtain and manage: a strong evidence of that is the effort in the development of tools able to manage this information (Pahl and Beitz, 2013; Ulrich, 2003). Companies frequently make use of Quality Functional Deployment (QFD) and requisites lists, users' needs, users' requirements with the purpose of tracking advantages (Carnevalli and Miguel, 2008). On the other hand, companies use Failure mode and effects analysis (FMEA) to gather and study drawbacks, failure modes and their effects and causes (Liu et al., 2013) However, product developers can acquire QFD and FMEA data only from the users of the invention: this leads to the need of expensive processes of customer's voice listening whose results are often unclear. Moreover, this information is not disclosed to researchers since this is part of the company know-how.

The description of the brought advantages and solved drawbacks are critical requirements for the patentability of a product, as stated by the politics and the guidelines given by World Intellectual Property Organization (WIPO) on writing patents (Organization, 2004). As stated by WIPO, an invention is a *solution* to a specific *problem* . The problem that an invention solves is a negative effect that state-of-the-art technologies can not fully overcome; on the other side, a solution is a way to solve this problem. A solution can lead to some advantages with respect to the known art. Thus, starting from the definition of invention, it is clear how it can be characterized by the advantages that it brings and by the problems that it solves. Also, it is reasonable to assume that having a clear picture of both advantages and drawbacks of a technology is important for an effective design process [6]

7.2.1 Methodology

In this section we show the approach used to extract the advantages and the drawbacks of the invention described in a patent. More precisely, the system identifies textual elements which represent the advantages or the drawbacks described in a patent. Advantages and drawbacks information can strongly advantage designers in the phase of new product development or in the process of product improvement and marketers in the phase of customer understanding and product placing. Furthermore a patent based method has a strong advantage with respected to methods that extracts information using other type of documents (e.g. online reviews of products (Mirtalaie et al., 2018a)): patents anticipate availability of products on the market by a factor varying from 6 to 18 months (Golzio, 2012).

The proposed extraction process is shown in figure 7.3 and its macro-phases are:

1. *Advantage and Drawback Clues Collection*: in this section we described the method to collect a reasonable number of generic advantage and drawback clues;
2. *Domain Clues Extraction*: in this section is shown how the generic clues are used in exploiting machine learning algorithm to extract new domain specific advantages and drawback clues;
3. *Domain Clues Validation* : since the new clues are automatically extracted, the output of the extraction phase surely contains a certain degree of noise. To clean this output a validation tool based on tweeter sentiment analysis is developed;
4. *Advantages and Drawbacks Extraction*: here the extracted clues (generic and domain specific) are expanded according to specific regular expression pattern in order to obtain the advantages and the drawbacks of the analyzed patent set.

[6]The most precise couple of words is **advantage** and **disadvantage**, but the reading is facilitated by using two very different words, therefore we decided to adopt the couple **advantage** and **drawback**.

Clues of Advantages and drawbacks in patents

First of all we have to define the concept of clue to an advantage or a drawback. To describe with a certain degree of precision an advantage or a drawback, patent writers need to use sequences of words of a certain length. Since NER systems do not perform well on long sequence of tokens, we split the problem of extracting advantages and drawbacks in two parts: first we extract entities that are clues in the sequence of words that describes the advantage or the drawback; then we extract the surrounding words to collect the whole sequence that describes the advantage or the drawback.

To better understand these concepts some examples are:

- ease of access
- cook food quickly and economically
- benefits of keeping an outdoor cooker lid fixed

For the present work, we refer to advantages and drawbacks as a sequence of words of minimal length that express the advantage or the drawback. The three phrases of the example are three advantages. On the other hand clues are words that are likely to be contained in advantages or drawbacks phrases.

Advantages and Drawbacks Clue Collection

The approaches to generate a knowledge base of clues were two. The first approach was based on a manual collection of clues of advantages and drawbacks directly from patent texts. This process was performed on 2,000 patents, randomly chosen from the freepatent database [7]. With this approach we were able to collect 3,254 advantages and 5,142 drawback clues. Some examples of the extracted clues are shown in table 7.7.

Table 7.7: Examples of the clues collected with the first approach.

Advantages Clues	*Drawbacks Clues*
ability	aggravated
efficacy	breakage
ensure	damage
healthy	defect
innovative	error
optimum	improper
protect	leak
quick-release	problem
reinforce	unavailable
securely	wrong

The second approach consisted in looking for alternative methods to indicate advantages or drawbacks clues, looking defined word patterns. The most relevant are the negations of advantages to obtain drawbacks, and the negation of drawbacks to obtain advantages. Some examples of such constructions are shown in table 7.8.

Table 7.8: Examples of the clues collected with the second approach.

Advantages Clues	*Drawbacks Clues*
non damaged	out of
anti-corrosion	in need of comfort

[7] http://www.freepatentsonline.com/

Advantages Clues	*Drawbacks Clues*
loss reduction	non user-friendly
prevent	diminish comfort
defect free	issue with
reduce waste	problem with
reduction of	lacks of
cost less	lacks with
less severe	loss of
avoid disease	unfacilitate

At the end of this process, a total of 6.568 advantages and the 14.809 drawbacks formed the knowledge base for the system, and gave us a reasonable number of clues to be used in the next step of the process.

The first approach was restricted by the lists being extracted from a random and limited sample of patents. On the other side, the rules used in the second approach are non exhaustive, and this can create non-sense clues, due to all of the possible combinations of words. Anyway, it is reasonable to assume that a large set of non-domain dependent clues are collected and will be used in the next steps of the process.

It is important to underline that the list of advantages and drawbacks clues is designed to avoid linguistic ambiguities when projecting these entities on the corpus. For example *guide* has two different meanings[8] when used as a verb or as a noun: as a verb it means *"to assist something or someone to travel through, or reach a destination in, an unfamiliar area, as by accompanying or giving directions"*, so it could be the clue to an advantage; at the same time as a noun it assumes the mean of *"a book, pamphlet, etc., giving information, instructions, or advice; handbook"* thus indicating a product and not an advantage. Avoiding such ambiguities is a crucial aspect to produce an informative training set, so ambiguous words were avoided to be projected on patents.

Domain Clues Extraction

The domain-specific clues collection process takes in input the automatically annotated patent set 1. The analyzed patent set is automatically annotated using the domain-independent clues, and is used to extract new domain-specific clues. We decided for the present methods to analyze the whole text of the patents and not to focus only on a specific section.

Our system resorts to state-of-the-art NLP tools which are part of the linguistic analysis pipeline shown in figure 7.5. In addition we developed a specific advantages and drawbacks clues extraction tool, still based on Natural Language Processing techniques.

The automatic patent set annotation 2 process, as shown in figure 7.5, is composed by a set of sequential steps. The first three steps are related to the linguistic annotation: sentence splitting and tokenization, part of speech tagging and lemmatization. Once these three steps are completed the entity extractor collects the advantages and drawbacks clues from the analyzed patents.

Sentence splitting and Tokenization steps split the text into sentences and then segment each sentence in orthographic units called tokens.

The *Part-Of-Speech tagging* (or POS tagging) step assigns unambiguous grammatical categories to the tokens. For the present application we use the most recent version of the Felice-POS-tagger described in (Dell'Orletta, 2009). Once the computation of the POS-tagged text is completed, the text is automatically *lemmatized* in order to group inflected forms of a word in a single item. Some of the following steps of the entire extraction process exploit the lemmatized texts in order to achieve better extraction results.

Successively the *semi-automatic annotation of advantages and drawbacks clues* is performed. The advantages and drawbacks clues extraction tool is the key ingredient of the present paper, and it is based on supervised methods. Such methods require an entity annotated corpus in order to extract new entities from unseen

[8] www.dictionary.com

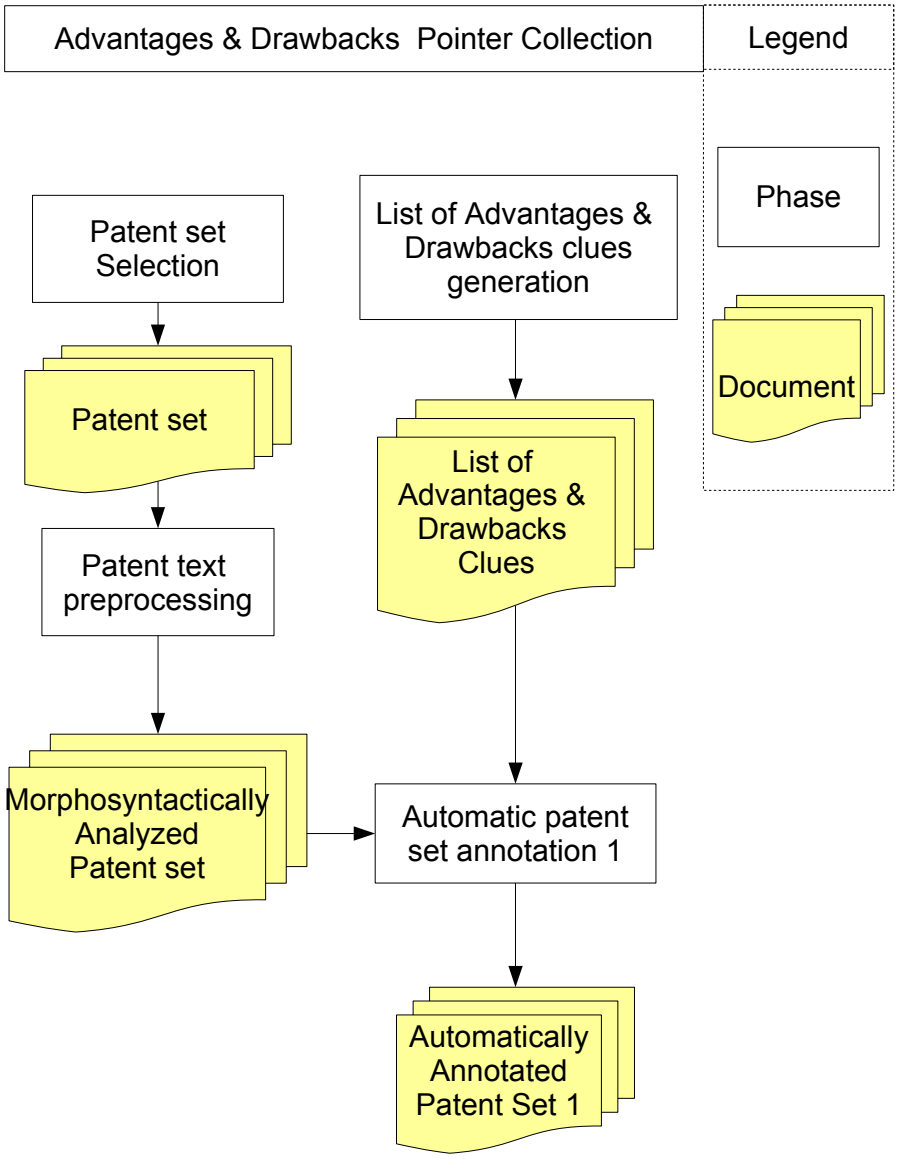

Figure 7.4: Main overview of the patent set annotation process.

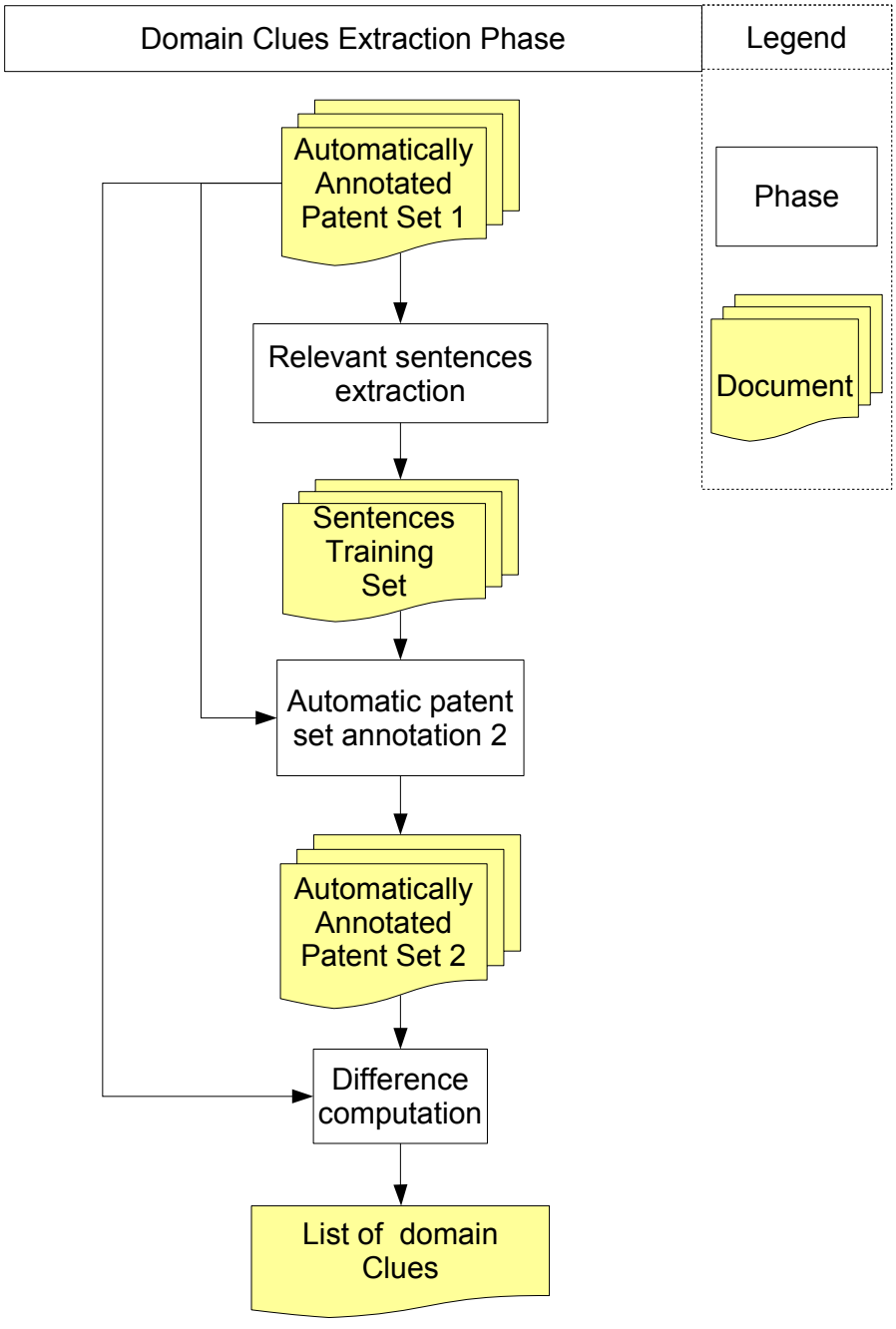

Figure 7.5: Overview of the domain specific advantages and failures clues extraction process.

documents. Since the manual annotation of a patent set is too expensive both in terms of time and manual effort, we apply a semi-automatic method to generate an advantage and drawback annotated corpus.

The entity annotation schema for a single token is defined using a widely accepted BIO annotation scheme (Ramshaw and Marcus, 1999):

- **B-ADV**: the token is the start of an entity representing an advantage clue;
- **I-ADV**: the token is the continuation of a sequence of tokens representing an advantage clue;
- **B-DRW**: the token is the start of an entity representing a drawback clue;
- **I-DRW**: the token is the continuation of a sequence of tokens representing a drawback clue;
- **O**: for the remaining case.

The *Advantages and Drawbacks Clues Extractor* is a supervised classifier that, given an annotated patent set, is trained on these examples. The patent set is: (a) linguistically-annotated, using the steps described above; (b) entity-annotated, exploiting the semiautomatic annotation process executed in the previous steps. Given a set of features the classifier trains a statistical model using the feature statistics extracted from the corpus. This trained model is then employed in the classification of unseen patents: it extracts new domain specific clues from patents and assigns them a probability score whether they are an advantage or a drawback. In our experiments the classifier has been trained using the Support Vector Machines (SVM) learning algorithm using the LIBSVM (Hearst et al., 1998a) library configured to use a linear kernel. The classifier uses two different kinds of features that are extracted from the text:

- **raw features**: prefix and suffix of the analyzed token; it works particularly well with advantages ending with -full -ious and with drawbacks starting with un- dis- etc..
- **word2vec features**: vector representations of words computed by the *word2vec* (Mikolov et al., 2013) tool.

Table 7.9 reports the detailed features chosen for the proposed advantage and drawbacks clues extractor.

Table 7.9: Context windows of the extracted features considering 0 as the current analyzed token.

Feature group	*Context Window*
Prefixes up to 4	0
Suffixes up to 4	0
Word2vec	$-2, -1, 0, 1, 2$
TAG	-1

By introducing prefixes and suffixes of the analyzed token, the classifier is able to identify frequent orthographic patterns which allow to maximize the precision in classification phase. On the other hand, the *word2vec* features are introduced in order to maximize the recall, since semantically similar clues should have similar *word2vec* vectors. Finally, the tag of the previous token is added to the final feature vector in order to improve the accuracy classification of multi-word clues.

Word2vec feature computation

While contextual, linguistic and compositional features are commonly used for entity extraction task in patents, from a computational linguistic point of view the presented system introduces the novelty of using *word2vec* features for entity extraction in patents.

Word2vec is a NLP tool able to produce word representations exploiting big corpora. The main property of the vectors produced by *word2vec* is that words that share similar contexts have similar vector representations. By using word vectors instead of the corresponding words we were able to overcome the problem of the limited lexical knowledge in the training phase.

To build our *word2vec* vectors we used the Skipgram model with a context window of 5 tokens. As reported in table 7.10, we used a corpus consisting of 48,194 different patents, containing more than 400,000,000 tokens.

The corpus was designed to contain patents belonging to different classes (12 in total) in order to acquire an extended knowledge of the contexts in which the words in general are surrounded. In addition, patents belonging to two of these classes are analyzed and, in the same section, detailed configurations of the entity extractor has been provided.

Table 7.10: Statistics of the documents on which the word2vec vectors have been learned. The patent sets of the analyzed case study are reported in bold.

Patent class	# Patents	# Tokens
A47G33	**2423**	**5.225.000**
A61G13	**2991**	**15.937.000**
A61G1	**5040**	**36.348.000**
A61H	**5199**	**41.831.000**
A61P25/24	5297	103.098.000
A63F1	5461	75.900.000
A63F3	4923	40.909.000
A63F7	4747	13.807.000
E02B3	3796	14.434.000
E04H9	2221	12.500.000
G01V11	1345	11.166.000
G08B13	4831	40.904.000
Total	48194	412.065.000

Clue validation using tweets sentiment analysis

Figure 7.6 gives an overview of the activities performed to validate the collected clues using twtitter. Since the manual review of the new domain specific clues can be very time consuming, an innovative approach to automatic validation of these entities is proposed. The approach is based on the assumption that advantages of technological innovations can be considered positive factors by the users. Conversely, the drawbacks of the artifacts are considered negative factors impacting on the satisfaction of the users. Both advantages and drawbacks are common terms or chunks of terms commonly used in other contexts, too. Therefore, if we can identify a wide source of sentences tagged with a polarity score and containing advantages or drawbacks, the probability of assigning the proper polarity to advantages and drawbacks increases.

Social media platforms provide powerful venues for consumers to interact not only with brands but also with other consumers as they engage in the processes of curation, creation, and collaboration (Evans et al., 2010). Such virtual platforms are places where users discuss about products, about their features but also about problems and failures they experienced during the daily use. The way they discuss or describe products or services is often unambiguous and highly polarized.

Our approach to the automatic validation of advantages and drawbacks exploits the information contained in the Twitter platform [9]. More precisely, for each extracted advantage or drawback clue we collect a set of tweets in which the clue is mentioned. Once a significant number of tweets is collected (in our case 3,073,959 - around 2,738 per entity in average), they are analyzed by a sentiment classifier. The main idea behind this process is to assign to each clue a sentiment polarity score which should express the feeling of the user with respect to the considered clue on the social media.

The tweet collection can be easily performed by using the Twitter streaming API [10], which is freely available.

[9] https://twitter.com

[10] https://dev.twitter.com/streaming/public

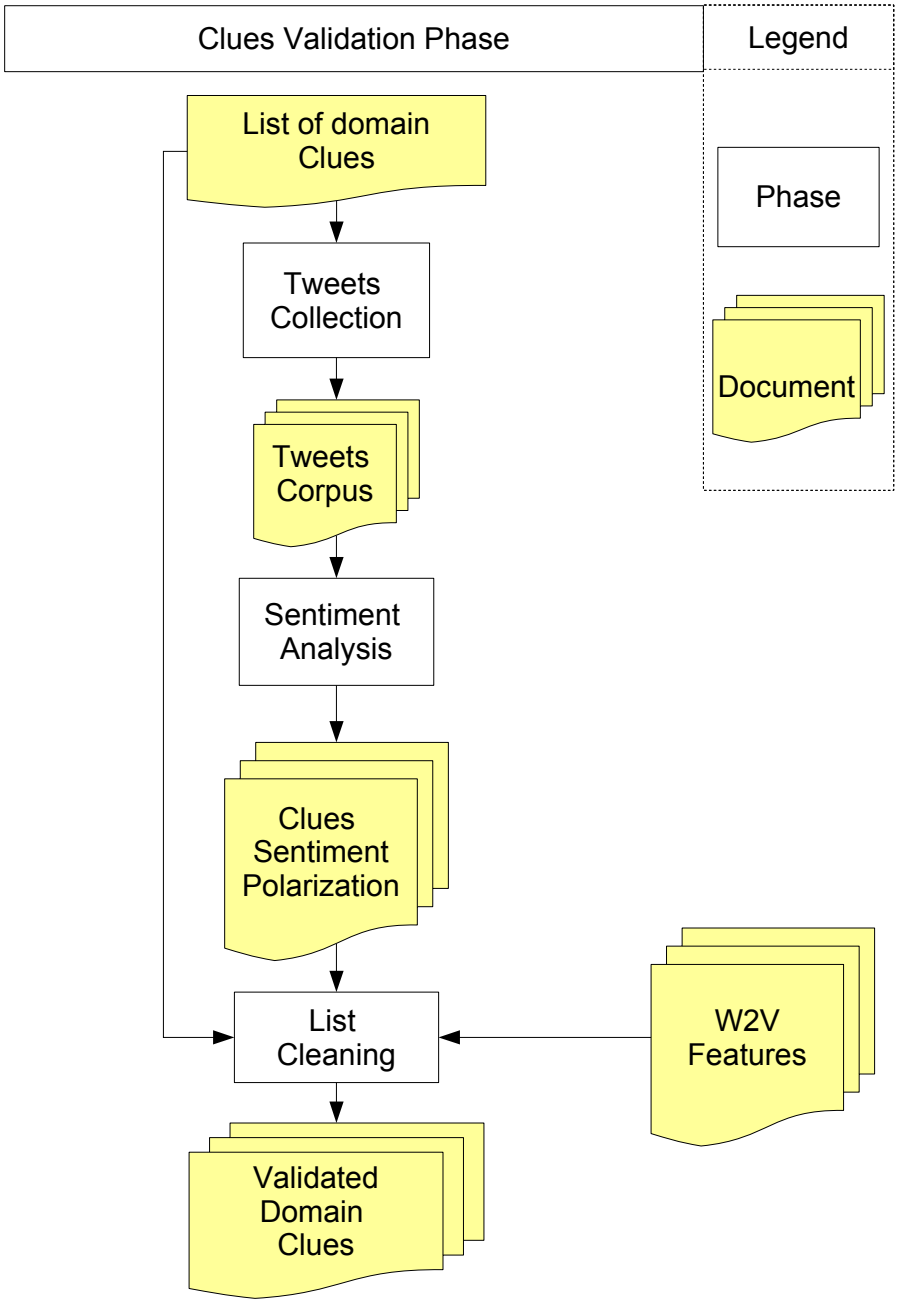

Figure 7.6: Overview of the domain specific advantages and failures clue validation process.

By assigning a polarity score to each clue, we expect to detect tagging anomalies: entities tagged as advantages by the classifier are expected to have a positive polarity in the extracted tweets. Vice versa, entities tagged as drawbacks by the classifier are expected to have a negative polarity in the extracted tweets.

Sentiment Classifier: features, classification model and performance evaluation

In our sentiment classifier we focused on a wide set of features ranging across different levels of linguistic description. The whole set of features we started with is described below, organized into four main categories:

- raw and lexical text features
- morpho-syntactic features
- syntactic features
- lexicon features.

This proposed four-fold partition closely follows the different levels of linguistic analysis which is automatically carried out on the text being evaluated, (i.e. tokenization, lemmatization, morpho-syntactic tagging and dependency parsing) and the use of external lexical resources.

Raw and lexical text features are extracted considering the text available in the tweet. For this work we considered a number of tokens, character n-grams, word n-grams, lemma n-grams, char repetition sequences, mentions number, hashtags number and punctuation.

Morpho-syntactic and Syntactic Features consider the part of speech tags and the syntactic analysis of the text. Our sentiment analyzer extracts Part-Of-Speech n-grams (coarse and fine), coarse grained Part-Of-Speech distribution and syntactic dependency types n-grams.

To extract features based on lexicons we exploited three freely available resources. The Bing Liu Lexicon (Hu and Liu, 2004), which includes approximately 6,000 English words, the Multi–Perspective Question Answering Subjectivity Lexicon (Wilson et al., 2005), which consists of approximately 8,200 English words and the SentiWordNet 3.0 Lexicon (Baccianella et al., 2010) which consists of more than 117,000 words. For each word in these lexicons the associated polarity is provided. In addition, we manually developed a lexicon of positive and negative emoticons, which usually is a strong indicator of tweets polarity. By exploiting the described resources, the following features were extracted: positive/negative emoticon distribution, sentiment polarity n-grams, sentiment polarity modifiers, the distribution of sentiment polarity, the most frequent sentiment polarity and changes of polarity in tweet sections. A more detailed description of these features is provided in [Cimino et al. (2014).

In order to assign a sentiment polarity score to each tweet, we employed an adapted version of the ItaliaNLP Sentiment Polarity Classifier for the English language (Cimino et al., 2014). This classifier operates on morpho-syntactically tagged and dependency parsed texts and assigns to each document a score expressing its probability of belonging to a given polarity class. The highest score represents the most probable class. Given a set of features and a training corpus, the classifier creates a statistical model using the feature statistics extracted from the training corpus. This model is used in the classification of unseen documents. The set of features and the machine learning algorithm can be parametrized through a configuration file. For this work, we used a tandem Long Short Term Memory Recurrent Neural Network (LSTM) - Support Vector Machines (SVM) architecture.

Validation of the extracted clues

The sentiment classifier is employed to validate the advantage and drawback clues which were previously extracted from patents by the clue extractor. In order to do so, we exploited the output of the tweet classifier on the tweets we previously downloaded. Each tweet, as said before, contains one or more clues to be validated. The sentiment classifier assigns the likeliness to each tweet to positive, neutral or negative. Consequently by analyzing all the tweets we previously downloaded, we obtained for each clue the distribution of positive, neutral and negative tweets.

Then to make a decision regarding each clue we used another SVM based classifier. This classifier is trained on a gold set of advantages and drawbacks clues: we manually labeled 344 words as advantages and 193 words as drawbacks, obtaining the gold set. The features used by this classifier are a number of positive, neutral and negative tweets extracted in which the entities are mentioned and the *word2vec* vector representing the considered entity. In table 7.11 the classification results of the proposed approach over a 5-fold cross validation are reported. The obtained results show that the proposed entity validation method is suitable for the automatic advantages/drawbacks clues validation process.

Table 7.11: Classification results of the proposed validation method over a 5-fold validation.

Method	Global accuracy	*ADV Prec.*	*ADV Rec.*	*ADV F1*	*DRW Prec.*	*DRW Rec.*	*DRW F1*
SVM-W2V	87.71	89.89	91.29	90.57	83.35	80.66	81.92

Advantages and Drawbacks sentences extraction

The advantages and drawbacks extraction process is shown in figure 7.7. Once all the domain specific advantages and drawbacks clues are extracted, these are merged with the ones belonging to the original knowledge base, obtaining a final list which will be processed by the advantages and drawbacks sentences extractor. The advantages and drawbacks sentences extractor exploits predefined linguistic and clues filters which operate on the automatic pos–tagged patents. Specifically, for each advantage and drawback term identified in patents, we used a pos–clue-pattern constraining the start-token and the rest of the token pos. Since we were interested in phrases containing words belonging to specific morphological categories, we identified sequences of allowed pos–clue-pattern in order to cover most of the English morphosyntactic multi–words structures, using the following pattern:

(ADVClue|textbar DISClue)+Noun.*Noun.*Noun.

The pattern is applied to the previously lemmatized text in order to have less sparse and more informative extractions.

This choice was made because the pattern:

1. expresses an advantage or a drawback exhaustively;
2. increases the precision and the recall of the final output list of advantages and drawbacks;
3. allows to build a three-level named based tree over the final output list.

In particular, the tree is built by grouping terms which share at the first level the same clue, at the second level the same noun and at the third level the same noun. This grouping procedure allows to easily represent the final output list in a tree structure which can be easily navigated by the end user of the system.

7.2.2 Results

In this section we describe the experimental use of the proposed process by applying it on four different patent sets.

To test the proposed methodology, we chose 4 patent sets composed of a sample of 3,000 patents each. The patent sets belong to 4 different IPC patent classes. The chosen classes and the definitions given by WIPO are reported in table 7.12.

Table 7.12: The patent IPC classes from which samples of 3,000 patents were chosen for the experimental analysis.

IPC name	*Definition*
A61G13	Operating tables and auxiliary appliances therefor

IPC name	*Definition*
A61H	Physical therapy apparatus
A61C15	Devices for cleaning between the teeth
A47J37	Baking; Roasting; Grilling; Frying

Our choice of patent sets aimed at challenging our system to find new domain specific advantages and drawback clues in different domains. Furthermore, we only selected patent sets from the IPC class A, which is based on human necessities, to maximize the probability of finding advantages and drawbacks that impacts on the users and not only on other products/components.

Once the advantages and drawbacks are extracted, a manual review process was performed on the output of the system to compute the number of true positive clues. In this way we were able to compute the precision of the process for both the advantages and the drawbacks. The output of the clue extraction validated by the clue validator is analyzed in table 7.13.

The table shows that the number of the extracted true positive advantage clues is higher than the number of the extracted true positive drawback clues. On the other side, the automatic evaluation process has a lower performance on advantage clues in terms of precision.

A first hypotheses to explain these results is that our knowledge base contained more drawback clues than advantage clues. Another possible reason could be that the applicant is minded to describe the invention highlighting the positive effects of the invention.

Table 7.13: Number of clues filtered with the clue validator and number of true positive clues.

	# Advantage clues	*# Drawbacks clues*
Tot extracted clues	3607	1244
Automatically Validated clues	1976	576
True Positive	984	448
Precision	49.8%	77.8%

In order to assess the performance of the overall process, another important measure is the amount of new information that is obtained, which we call *information gain*. As shown in table 7.14, the percentage of new discovered clues decreases with the number of starting clues. Obviously, the more patent sets are analyzed, the less new generic and domain specific clues are extracted. The percentage of information gain (represented as delta in the table), stabilizes at a 5% value in the advantage clues case, and 1% in the drawback clues case. This trend could be an evidence that the clue extraction process has a natural saturation level.

Table 7.14: Information gained by applying the extraction process on different patent sets. Each row reports the percentage of information gained by incrementally adding the extracted entities to the knowledge base and the overall number of entities belonging to the extended knowledge base.

Patent set	Adv.	# Adv.	Draw.	# Draw.
Knowledge Base	N/A	6,568	N/A	14,809
A47J33	+23%	8,133	+3%	15,332
A61C15	+12%	9,178	+2%	15,644
A61G13	+5%	9,653	+1%	15,849
A61H	+5%	10,175	+1%	16,053

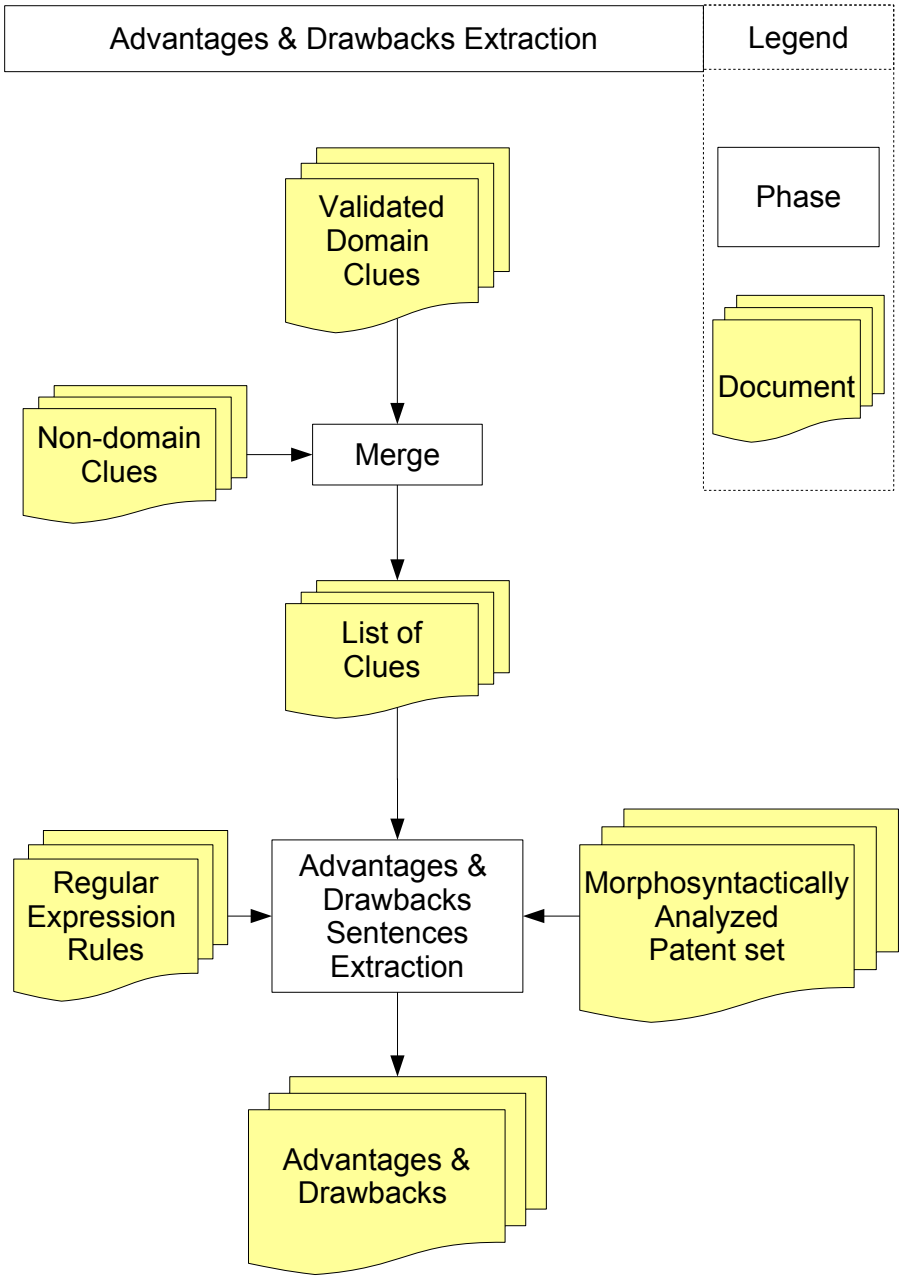

Figure 7.7: Overview of the advantages and failures extraction process.

Tables 7.15 and 7.16 show the frequencies of a randomly chosen set of the new extracted advantages and drawbacks domain specific clues for each of the four analyzed patent sets. The results show that the domain specific clues clearly characterize the different technical areas of the patent sets. It is thus interesting to notice how valuable information is contained in the context specific clues them self. Further research can decide to stop here the process, without extracting the whole sentence.

Table 7.15: Extracted domain specific advantages clues with the measures of occurrences for each patent set.

A47J37 (Baking)	A61H (Therapy apparatus)	A61C15 (Teeth cleaning)	A61G13 (Operating tables)
transport 295	regenerative 144	elasticity 495	rigidity 784
integrity 246	waterproof 101	rigidity 461	ventilation 177
rigidity 233	hygienic 85	disinfection 247	hygiene 135
insure 180	ergonomically 77	precision 199	versatility 121
adjusting 164	disinfection 48	ergonomically 108	reliably 113
unobstructed 73	prevent excessive 39	economically 105	disinfection 56
uniformity 64	hemodynamics 25	waterproof 81	humidification 48
sensitivity 44	prophylaxis 22	hygienically 33	ergonomically 39
hygienic 30	prevent slippage 21	quick-connect 26	sanitation 20
selectively 34	smoothly 15	sanitation 24	non-invasive 12

Table 7.16: Extracted domain specific drawbacks clues with the measures of occurrences for each patent set.

A47J37 (Baking)	A61H (Therapy apparatus)	A61C15 (Teeth cleaning)	A61G13 (Operating tables)
accidental 61	infection 595	infection 446	syndrome 134
burnt 59	trauma 378	inconvenience 126	costly 72
malfunctioning 12	abrasion 106	irregularity 40	claustrophobia 38
time-consuming 8	fragmentation 37	pathogen 14	malfunctioning 36
non-compliant 6	paralysis 17	infected 8	unnecessarily 27
dirty 6	hematoma 15	unintentionally 6	discoloration 19
ignites 4	uncomfortably 8	abnormal 6	hyperglycemia 11
turbulence 4	undetectable 8	burn 4	unavoidable 10
cross-contamination 3	embarrassment 8	toxic 3	not linearly 8
violently 3	discoloration 8	erosive 3	catastrophic 6

Then, after the re-projection of the extracted clues on the text, the regular expression described in section is used to extract the sentences highlighted by the clues.

The number of advantages and drawbacks sentences extracted from each patent are shown in table 7.17. The table shows that the occurrence of sentences describing an advantage is higher than the ones containing a drawback 7.14. This result may be due to the fact that the applicant is minded to describe the invention by highlighting the positive effects of the invention.

Table 7.17: Number of sentences containing advantages and drawbacks for each analyzed patent set.

Patent class	*# Advantage sentences*	*# Drawbacks sentences*
A61G13	7,836	1,048

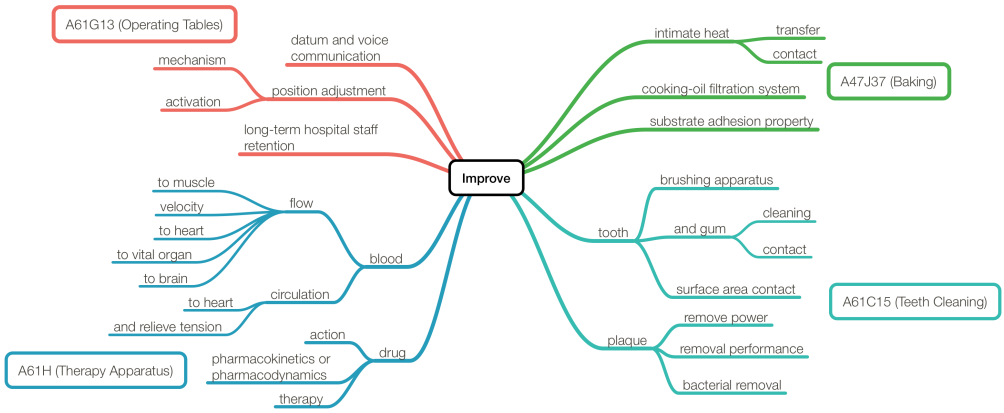

Figure 7.8: Sample of the tree based taxonomy extracted from the analyzed patent sets. The sample contains some of the leaves linked to the advantage clue Improve

Patent class	*# Advantage sentences*	*# Drawbacks sentences*
A61H	10,879	1,463
A61C15	9,551	1,572
A47J37	9,973	1,662
Total	38,239	5,745

Figure 7.8 and 7.9 show two subsets of the taxonomies obtained by the extraction of the advantages and drawbacks for each of the four analyzed patent sets. The two figures respectively refers to a subset of the leaves linked to the advantage clue *Improve* and a subset of the leaves linked to the drawback clue _Damage. In both cases an additional trimming action is performed by removing those branches or leaves containing terms belonging to the stop-word list typical of patent lexicon (e.g. claim, embodiment, invention, comprise, figure, etc..). The figure shows that our process can extract highly informative sentences also starting from generic and non-contextual clues like *improve* or *damage*. Moreover the words that follow the generic clues are specific of the technical field of the analyzed products. Both the results are promising for future applications, especially for the design fields. In particular figure 7.8 allows designers to focus on the positive side of the effects provided by the product and to better meet the explicit and implicit user needs. Similarly, figure 7.9 helps designers to redesign of the product in a proactive way, to keep attention to the critical issues identified by the drawbacks and to conceive possible corrective actions to solve such drawbacks.

7.3 Trademakrs

The market interest for both patents and trademarks has increased during the last decades, with a significant increase in filing for both types of Intellectual Property. The academia also took attention to patent data and trademark data, but with different and (almost) disconnected research approaches. Furthermore, the analysis of patents to study R&D is predominant while less attention has been paid to trademarks (Griliches, 1981). Indeed, very few are the works where patents and trademarks are investigated together and not generically cited together as parts of the intellectual property right framework. Moreover cultural differences exist between United States and Europe (at least) and practical consequences can be observed in daily life: for example, even if Europe had an earlier and more enduring interest in trademarking than the US the presence of the symbols ® and ™ in the product labels are more common in US than in Europe (MERCER et al., 2010). The difference exists also from an industrial perspective, across different sectors (Baroncelli et al., 2004). Indeed, trademarks have their larger use worldwide in the R&D intensive scientific equipment, pharmaceuticals sector and advertising intensive manufacturing industries (clothing, footwear, detergents and food products).

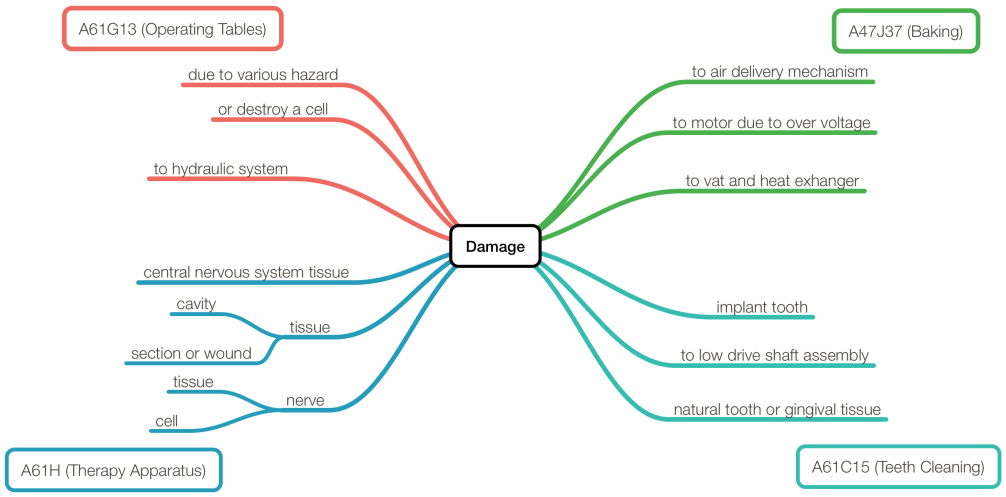

Figure 7.9: Sample of the tree based taxonomy extracted from the analyzed patent sets. The sample contains some of the leaves linked to the drawback clue Damage.

The present chapter studies the usage (if any) of *trademarks* and *trade symbols* within patent texts. The aim is to investigate possible indicators about IP strategies and innovation output utilizing information about the interlink between patents and trademarks. An additional goal is to understand if patent applicants makes a right use of trademarks and trade symbols when writing a patent document. The analysis clearly stands at the boundary between patent and trademark research areas, and since such intermediate territory has not been investigated in depth, there are several research questions we would like to answer. The starting point is a sheer numerical evidence: we have found out that, at the date of 01/09/2017, a total of 2.162.962 patents worldwide (for details about the database and its coverage see section 3) contain a trademark symbol. It is therefore a not negligible phenomenon, worth of investigation and that should provide interesting insights about the mechanisms and strategies that companies adopt at the frontier between technical innovation and marketing.

The use of trademark symbols ™ and ® in patents from a legal perspective

It is evident from our preliminary analysis that there exist an high number of patent documents containing a trademark indication. In the present section we investigate what is the legal perspective of this usage.

In (Butler, 1969) the authors published a series of rules to standardize the use of trade terms in patents. They also specifies different rules in the case of use of trade terms in specifications and the use of trade terms in patent claims. The rules are:

1- A trade term is properly used in a specification if those skilled in the art can make the product designated by the trade term at the time the application is filed, using the specification and/or published literature that is implicated by the specification. 2- A trade term is also properly used in a specification if the product is generally known to persons skilled in the art and is readily obtainable at the time the application is filed, provided the composition of the product is a trade secret and there is reason to believe that whenever the composition of the product is modified the trade term will also be changed. 3- A trade term is also properly used in a specification if it designates a component of the embodiment which is not essential to the invention. 4- A trade term can be used in a claim only if its meaning has been adequately defined in the specifications, whereby it imparts specific limitations to the claim."

The second rule is particularly interesting because it seems preventing early patenting or patenting before having produced and used a trade name. Rule number two is even more interesting since it contains a case of a product that is a trade secret used to build another product subject to patent application.

Similar guidelines are contained in the European patent convention (Hall and Helmers, 2018). Here is state

that is not desirable to use trademarks or trade names if such words merely denote origin or where they may relate to a range of different products. Anyway trademarks could be used in patent application to satisfy art.83, that states that the application shall disclose the invention in a manner sufficiently clear and complete for it to be carried out by a person skilled in the art. In this case the product must be sufficiently identified, without reliance upon the word. A special case are such words that have become internationally accepted as standard descriptive terms and have acquired a precise meaning (e.g. “Bowden” cable, “Belleville” washer, “Panhard” rod, “caterpillar” belt). In this case they may be allowed without further identification of the product to which they relate. It is clear specifications afflicts the analysis that we want to carry out. Always in the same document are given also guidelines for the usage of trademarks in claims. These are, as expected, different from the specifications. For claims we have the problem that it may not be guaranteed that the product or feature referred to is not modified while maintaining its name during the term of the patent. They may be allowed exceptionally if their use is unavoidable and they are generally recognised as having a precise meaning. It is the applicant’s responsibility then to ensure that registered trademarks are acknowledged as such in the description. From that we can deduce that the presence of a trademark in patents and more specifically in claims decreases the reproducibility of the invention and thus the quality of the patent. Also the the US patent legislation (Jaffe, 2000) focuses on the use of trademarks in patents claims. Here the presence of a trademark or trade name in a claim is not considered improper but the examiner should analyze the claim to determine how the mark or name is used. In fact, the trademark or trade name should identify a source of goods, and not the goods themselves. In this case the claim scope is uncertain since the trademark or trade name cannot be used properly to identify any particular material or product. In fact, the value of a trademark would be lost to the extent that it became descriptive of a product, rather than used as an identification of a source or origin of a product. Thus, the use of a trademark or trade name in a claim to identify or describe a material or product would not only render a claim indefinite, but would also constitute an improper use of the trademark or trade name.

Finally, in (Pressman and Stim, 2018) the authors gives some guidelines on how to introduce trademarks when it is unavoidable. First, the trademark should be capitalized and used as an adjective (not a noun), followed by the generic name of the product or service. Furthermore when referring to the trademark there should also be a reference to the trademark owner.

7.3.1 Methodology

With respect to Users 7.1 and to Advantages and Distadvantages @ref(#advdrwresults) the process to extract tradenames from patents is trivial. The main activity performed are (for details aobut the activities see section 5:

- Sentence splitting
- Tokenization
- Part-of-speech tagging
- Lemmatization

After that is is possibile to identify tradenames searching for the symbols ™ and ®.

7.3.2 Results

The first investigation has the aim of understanding if the ™ and ® symbols has different content depending on different IPC (Organization, 1971) classes. The IPC classes are: The main evidence in figure 7.10 is that ® is used more than ™ in patent, with and average percent of presence of 7.4% and 5.6% respectively. The reason of this evidence could be that patents contains unregistered trademarks that are never converted to registered trademarks; another reason is the confusion between the ® and ™ symbol, usually considered as synonymous. Furthermore the writer will use the symbol only if he/she knows that the mark is protected, otherwise he/she will not indicate any symbols or will use ™ or ® without checking the right use. Another possible problem is that the word has become of common use and thus usually used without symbols.

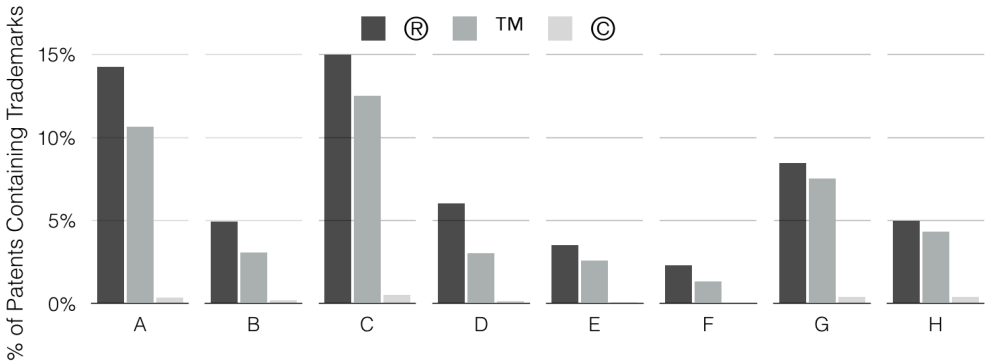

Figure 7.10: Histogram of the percentage of ® and ™ in patentes for each IPC class (limited to US patents). A (Human Necessities); B (Performing Operations; Transporting); C (Chemistry; Metallurgy); D (Textiles; Paper); E (Fixed Constructions); F (Mechanical Engineering; Lighting; Heating; Weapons; Blasting); G (Physics); H (Electricity).

Furthemore, classes A (Human Necessities) and C (Chemistry; Metallurgy) clearly has more trade symbols than the other classes.

Secondly, from table 7.3.2 it is evident that the average ratio is 6.4 for ® (the usage is 6.4 times higher in the year span 2007-2014 than the span 1995-2002) and 5.4 for ™ (the usage is 5.4 times higher in the year span 2007-2014 than the span 1995-2002) thus the usage of trade-symbols in patent is growing and the presence of ® is increasing faster than the presence of ™. From the present results is not possible to say if this effect is due to an higher quality of the process of patent application (and thus an higher awareness of applicant and examiner) or of a positive trend in trademark registration. In particular for the IPC class E (fields constructions) the number of ® and ™ has increased 10.5 and 9.0 times respectively. This could be an evidence of the fact that the number of trademark in the field of fixed constructions is increasing.

Table: For each IPC class the ratio of the percentage of patents containing at least one ® or ™ character is computed for two patents sets 1995 to 2002 and from 2007 to 2014.

IPC Classes	Ratio ® 2007-2014 over ® 1995-2002	Ratio ™ 2007-2014 over ™ 1995-2002
A	5.9	5.0
B	6.2	5.4
C	5.9	5.2
D	5.6	5.7
E	10.5	9.0
F	7.5	6.2
G	4.6	3.2
H	5.1	3.2
Average	6.4	5.4

The Selected Tradenames

We a set of popular tradenames referring to (Morris, 2016). Then we filtered the ambigouus ones (the one that can refer to surnames), because this can create an onverstimation of the number of patent citing it witouth the tradenames. The total namber of tradenames is 38 and these are:

- amazon®
- blackberry®
- bose®
- budweiser®

- chiquita®
- chrome®
- coca-cola®
- ebay®
- facebook®
- fender®
- firefox®
- gibson®
- gillette®
- heineken®
- ibanez®
- intel®
- iphone®
- kellog®
- kodak®
- lego®
- marlboro®
- matlab®
- mcdonald's®
- nitinol®
- nutella®
- nylon®
- photoshop®
- polaroid®
- post-it®
- powerpoint®
- prozac®
- sennheiser®
- spotify®
- teflon®
- velcro®
- whatsapp®
- xanax®

In figure 7.11 are shown the number of patent containing the tradenames with and withouht ™ and ®; furthermore the information is divided for all the years, from 1995 to 2002 and from 2007 to 2014. In the previous analysis we noticed that ® is generally more used than ™. Now the goal is to understand if patent writers uses tradesymbols or not when citing tradenames, if there are any differences between different tradenames and if we can notice different behaviours in recent years. From figure 7.11 is evident that despite the fact that it is mandatory to use ™ o ®, these symbols are not always used.

Tradenames Usage in Patents

It is relevant to understand if there exists any difference between different tradenames in the correct usage of trade symbols (a trade name has been cited in the patent texts with the trade symbol). Figure 7.12 illustrates in a bi-logaritmic plot the distribution of trade names correctly cited in patents with ® and ™ (Y-axis) versus the same trade names cited without using the proper symbol. Most of trade names are located under the diagonal. Those around the diagonal are those we can considered as well known trademarks and not so subject to genericization phenomena (green area). Conversely on the right we can find a read area where the ratio is even more shifted toward the absence of any indication of protected trademark. In the middle the orange area where many of the most famous trademarks fall down and that could be in phase of generalisation or conversely, since the trademark of a product is totally entangled with the owner (Iphone-Apple; Intel), inventors do not feel the reason to cite them as a trademark. A remark on Figure 7.12 is necessary: the plot presents a correct measurement of trademarks with ® and ™ plotted in the y axis, while

Trademark	All			1995-2002			2007-2014		
	TradeName Total	TradeName **with** ™ OR ®	TradeName **without** ™ OR ®	TradeName Total	TradeName **with** ™ OR ®	TradeName **without** ™ OR ®	TradeName Total	TradeName **with** ™ OR ®	TradeName **without** ™ OR ®
nylon®	431859	3838	428021	70340	174	70166	147870	1915	145955
teflon®	199084	47875	151209	29083	2237	26846	69305	23873	45432
intel®	131103	25148	105955	21741	857	20884	51497	13748	37749
chrome®	97493	2523	94970	11918	0	11918	39172	1545	37627
velcro®	75657	24860	50797	9459	925	8534	27864	12643	15221
nitinol®	54813	3031	51782	5186	136	5050	27298	1646	25652
kodak®	54619	772	53847	13219	64	13155	15013	416	14597
iphone®	40009	17523	22486	26	2	24	28718	12480	16238
blackberry®	37904	13967	23937	587	56	531	25709	9883	15826
facebook®	32903	10959	21944	0	0	0	22671	7551	15120
matlab®	26771	6086	20685	1526	149	1377	15270	3680	11590
firefox®	19280	5051	14229	7	0	7	13655	3687	9968
fender®	19127	53	19074	2367	2	2365	6083	37	6046
gibson®	18717	46	18671	2601	1	2600	7668	28	7640
amazon®	15840	2284	13556	309	21	288	9624	1431	8193
photoshop®	14595	3436	11159	1695	237	1458	7182	1824	5358
mcdonald's®	12405	346	12059	1891	19	1872	4684	221	4463
powerpoint®	11065	2958	8107	866	149	717	5715	1561	4154
ebay®	8304	1758	6546	284	49	235	5087	1040	4047
post-it®	6689	1069	5620	896	70	826	2651	452	2199
polaroid®	6293	195	6098	1375	24	1351	1308	81	1227
bose®	5922	121	5801	462	0	462	3126	72	3054
prozac®	3788	1899	1889	320	66	254	1919	1010	909
coca-cola®	3178	878	2300	283	41	242	1679	510	1169
kellog®	2732	44	2688	0	0	0	801	27	774
gillette®	1799	77	1722	397	3	394	576	47	529
lego®	1597	700	897	168	16	152	755	360	395
marlboro®	1263	47	1216	335	0	335	384	23	361
xanax®	1202	680	522	92	3	89	638	433	205
spotify®	1142	258	884	0	0	0	622	115	507
ibanez®	694	11	683	113	0	113	295	8	287
whatsapp®	470	98	372	0	0	0	93	21	72
budweiser®	350	71	279	47	5	42	187	42	145
sennheiser®	211	6	205	47	0	47	76	3	73
heineken®	145	31	114	15	1	14	77	11	66
chiquita®	45	7	38	9	0	9	23	3	20
nutella®	41	23	18	1	0	1	26	16	10

Figure 7.11: Number of patents citing one of the 38 selected tradenames. The data are divided for any years, from 1995 to 2002 and from 2007 to 2014.

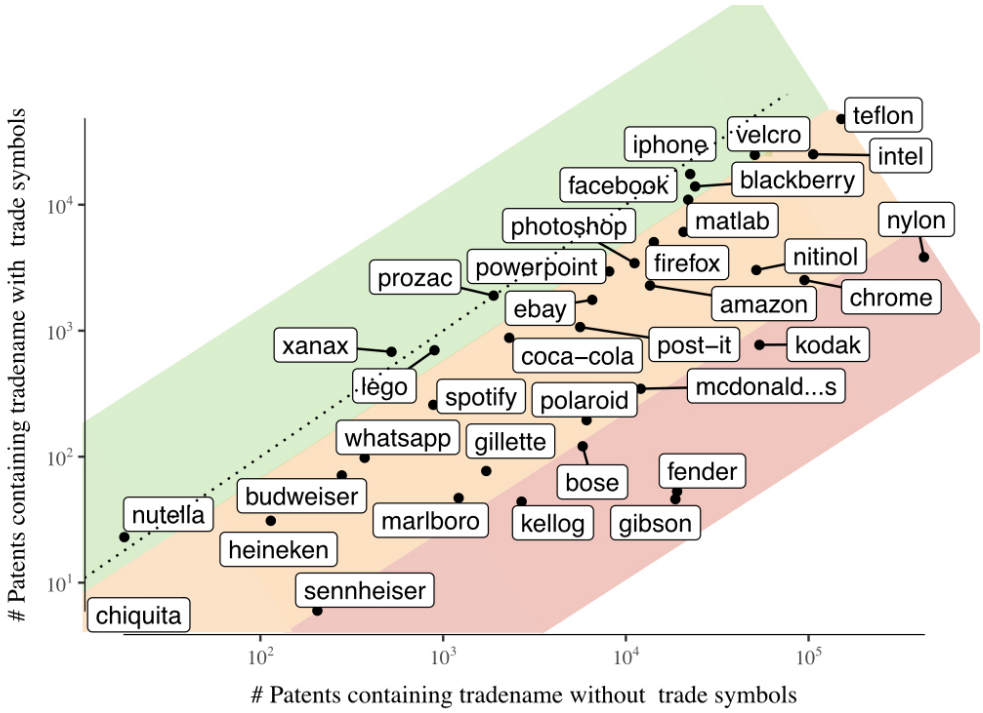

Figure 7.12: Plot of the trade names correctly cited in patents with ® and ™ on the y-axis and the same trade names cited without using the proper symbol on the x-axes. Both axes are on a logaritmic scale.

the search for trademark with missing trademark symbols in some case can introduce undesired errors and include other meaning of the word. Take for example the case of Blackberry®: the datum is referring to mobile phones and accessories or software applications. In all the cases where ® or ™ are correctly written, it is probably true that the inventor is referring to the Blackberry® mobile phones, but machines for jam or juice production extracted out of the blackberry introduce false positive examples. Similarly Fender is probably located too far on the right because it is also a part of a car, bike, motorbike, boat, etc.. and Gibson could introduce citations of works performed by someone called Gibson, therefore if cleaned results are necessary the searching strategies for the trademarks without symbols have to be refined.

The Popularity of Tradenames in Patens

In figure 7.13 are plotted the trends of the numbers of patents that correctly cites a tradenames. We anlyzed 12 different tradenames, divided in couples: each couple belong to a similar sector. We make use of a generalized additive model fitting function to better analyze the results and to make more easily to understand and compare the two element of each pair.

Each pair has been chosen to compare trademarks in homogeneous markets. Some of them shows:

- (d, e) similar trends but different incidence;
- (a, c) dissimilar behaviours (Amazon vs Ebay, Blackberry vs. Iphone: linear vs. exponential);
- (f) similar results (e.g. Gibson vs Fender in the music market have almost the same behaviour and a reduced presence in patents)
- similar behaviour but shifted in time (e.g. Firefox vs Chrome).

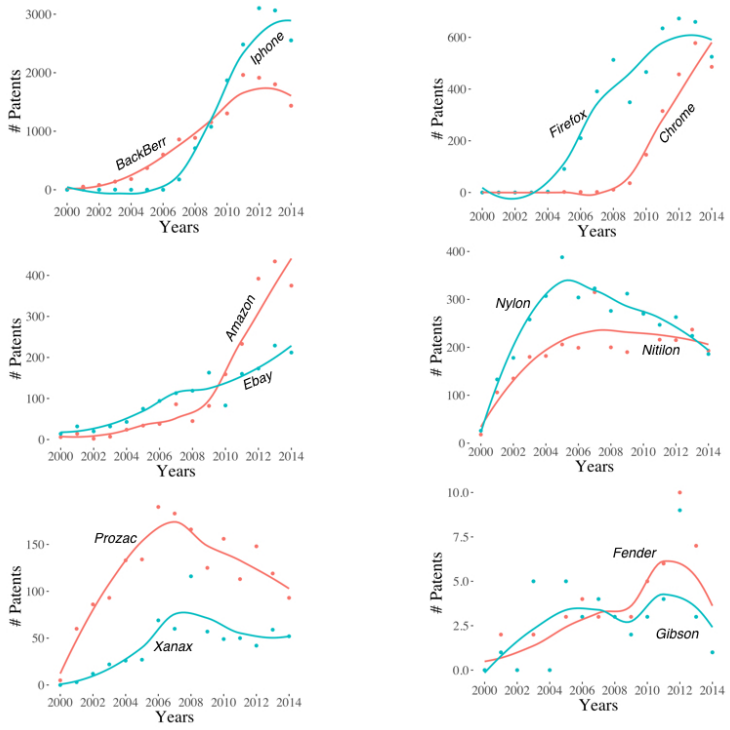

Figure 7.13: Plot of the trade names correctly cited in patents from 2000 to 2014.

The Similarity of Tradenames in Patens

Further informations can be derived by analysing the IPC classes where the patent have been chosen by the assignees-attorneys-examiners.The choice of a class is not only matter of market segment but rather of market use. Moreover it is necessary to remark that from a formal point of view the IPC classes form a feature vector, denoted y N639. IPC classes form a complete dataset of 639 elements of such feature vector. Each tradename is represented in such a feature vector where the number of patents in each class constitutes the length of the vector along that direction (feature). Our training data is therefore D = (x,y,z), where x [1, .., 12], y [1, .., 639], and z represents the patent numerosity z [0..15*106]. This dataset contains information about how z varies as a function of x and y. The matrix is quite empty since each product/brand addresses needs of specific market/s, after cleaning of the totally empty columns the matrix reduces to the dimensions of 12x479 and demonstrates a correct choice of the trademarks since they covered almost 75% of the IPC space.

With such a dataset a series of computations can be performed: here we show the results of the correlation analysis among the chosen trademarks to analyse their relative positioning on the invention landscape (not only the market); Correlation analysis is shown in Figure 7.14 The matrix is triangular owing to the symmetry of the relationship. The main evidences are the following:

Figure 7.14: Heat-map of the IPC-based similiraties between trademarks.

Chapter 8

Papers

There are fields (mature technologies) where patents anticipate papers, while in others (basic research) the opposite happens. For this reason, scientific literature is the place where companies and scholars gather information about the problems that researchers are facing in the development of new technologies. Anyway, standard approaches of knowledge extraction from papers requires skilled personnel, they are time consuming and lacks in reproducibility. Furthemore, the volume of scientific literature has grown rapidly raising an imminent question about how to extract knowledge from this source. On the other hand, this tehcnical knowledge if managed, can be used to answer questions that 10 years ago would need the help of domain experts to be answered. In the present section is shown how text mining techniques can help to unswer some of this questions. In particular, it is an essential problem when different classification codes are used in order to organize scientific knowledge on a specific domain, becaues a specific categorization in a certain scientific field is missing. This leads to unnecessary complications in the researchers' aims who want to quickly and easily find literature on a specific topic among the large amount of scientific publications, or want to effectivelly position a new research. Text mining techniques, and in particular topic modelling, can help scholars in solving this problems, if properly used togheter with domain expertise. A deeper description of what patents are and how these documents are used to mine technical knowledge can be found in section 6.2

In this section we present two methodologies capable of automatically segmenting a knowledge field. These selected knowledge field are: sustainable manufacturing and block-chain. The results of the methodologies are described, togheter with example of applications.

8.1 Sustainable Manufacturing: an Analysis of the 6R Framework

One of the most important objectives of Sustainable Manufacturing (SM) is developing innovative and viable engineered materials, manufacturing processes and systems to provide multiple life-cycle of products. In SM the old concept "from cradle to grave" is now transforming into "from cradle to cradle" (Jawahir and Bradley, 2016), tending toward multiple product life-cycles or even a "near-perpetual" product/material life. Scientific contributions in the sustainable manufacturing field mostly deals with energy and resource consumption. In this respect, two different main fields of causes can be identified: the process level and the material efficiency one. As a matter of facts, manufacturing processes have a significant role also in putting in place material efficiency strategies (Ingarao, 2017). As far as the processes are concerned, a first classification of research contributions was discussed in the CIRP General Assembly (Duflou et al., 2012). There the authors state that research in manufacturing field, oriented to environmental impact reduction, can be clustered in 5 main sub-classes:

1. unit process level (Individual device or machine tool in the manufacturing system)
2. manufacturing system level,
3. facility,

4. multi-factory system up to considering the whole
5. supply chain level.

Another review paper was presented at the ASME international manufacturing science and engineering conference (Haapala et al., 2013). In that paper the authors scrutinize the research papers focusing more on the differentiation between manufacturing processes and manufacturing system. Considering the process level, Ingarao (Ingarao, 2017) clustered the scientific papers in 4 subsections: 1. effect of process parameters 2. role of machine tool architecture and related technology 3. applied process 4. manufacturing approach selection.

As concerns material efficiency options, a framework is presented by (Allwood, 2014). The authors there provide strategies within three main classes corresponding to principles: reduce, reuse and recycling. Within each class guidelines for material efficiency practices are detailed. As concerns material efficiency options, a framework is presented by Allwood (Allwood, 2014). There the authors provide strategies within three main classes (i.e. principle): reduce, reuse and recycling. Within each classes guideline for material efficiency practices are detailed. A good framework of all the possible reuses of materials is provided by Coopper (Cooper and Allwood, 2012). In this research the authors identify four main reuse strategies for metals: Remanufacture, Reshape (applying metal shaping processes, additive, subtractive, mass conserving) to obtain a new geometry, Relocate: (recovering component and applying little refurbishment, components reused in the same type of products), Cascade: recovering component and use it in another less demanding use (downgrading). The role of manufacturing processes in putting in place material efficiency/reuse strategy was also outlined by Ingarao (Ingarao, 2017). In fact, manufacturing processes deserve to be considered as means for enabling material efficiency strategies. SM may be pursued by several strategies, such as re-designing and/or even changing manufacturing practices to conceive new-generation products, as well as by creating a closed-loop of environmentally-friendly material flow.

The 6R Framework

All the life-cycle stages of the product development (namely, design, production, use and post-use) should be considered carefully, with a particular attention to the design phase. This represents the real difference introduced by sustainable manufacturing concept with respect to traditional manufacturing, to lean manufacturing or even to green manufacturing, precursors of the SM. In a word, SM paradigm embraces those principles belonging to the 6R framework, namely: Reduce, Reuse, Recycle, Recover, Redesign and Remanufacture. The latter three principles where not included into the previous 3R framework. To a certain extent, it appears that the better definition of sustainability for any manufacturing operation provided so far is the extent to these principles are applied. This statement justifies because in the search to find a common rationale behind the mess of SM applications, one possible way is to find common roots in the principles that inspired the same applications: namely principles, which allows to the decision maker either an effective way to search for new sustainable solutions or to a certain extent measure the "degree of sustainability". The higher the number of principles that are satisfied, the higher the potential positive impact on sustainability can be for a given adopted solution. In this sense, far from stating that the 6R framework is a widespread and commonly adopted metric of SM, the starting assumption in this paper is that 6R is the only approach, to the author's knowledge, that provides a criteria of classification and selection of solution, and thus indirectly to provide a clear definition of what a SM application may look like. The true question nowadays is *if the 6R framework may capture the essence of SM*, provided that it is really critical to clearly define the SM paradigm rationale to the scientific community. This paper aims to stimulate the reflection from the Italian perspective to this point, with a particular concern to the production technology side, by using an automatic classification of papers within the 6R framework and then by benchmarking approaches followed by the Italian Technologist research-network SOSTNERE on SM issues.

Sustainability (from sustain plus ability) refers to the set of properties of a given system (either natural or artificial), which allows the same system to maintain itself for an almost indefinite period of time. This concept was officially introduced in a document of the World Commission on Environment and Development (WCED) entitled "Our Common Future", where the Sustainable Development was defined as follows: *"Sustainable development is development that meets the needs of the present without compromising the ability of*

future generations to meet their own needs"(WCED, 1987). Starting from this general definition, further conceptualizations have been produced for manufacturing activities and/or processes, as here briefly recalled. According to the Organization for Economic Co-operation and Development, Sustainable manufacturing is a formal name for an exciting new way of doing business and creating value. Different statements of sustainability in literature [1] share the same focus on the following three aspects: economy, environment and society. It is thus possible to summarize from the literature analyzed a possible definition for sustainable manufactured processes/products, according to the following set of prescriptions:

- (-) minimize business risk;
- (-) minimize negative environmental impacts;
- (-) conserve energy and natural resources;
- (-) are safe for employees, communities and consumers;

- (-) are economically sound;
- (+) a new way of creating value;
- (+) are socially and creatively rewarding for all working people;
- (+) providing access to basic services, green and decent jobs and a better quality of life for all ;
- (+) adopt sustainable infrastructures.

where the (-) sign stands for those prescriptions oriented to preservation of resources without any significant change of the present condition, while the (+) sign indicates those prescriptions aiming at ameliorating/modification of the trends with respect to traditionally manufactured processes/products. To summarize simplifying, sustainable manufacturing is all about minimizing business risks of any manufacturing operation while maximizing the new opportunities that arise from improving processes and products.
Fascinating in principle, these general statements are really difficult to deploy into real operations and production settings. The focus of the present analysis is to derive a clear definition of the concept of manufacturing sustainability based on clear and paper based evidences, rather than referring to general ethical or social principles, which appear to be a **top down** definition. Conversely, information extracted from all the scientific papers available on SM can provide a sort of **bottom up** statement. Evidences are relevant words and short concepts that can be related to sustainability, as it will be explained in the following paragraphs by also providing some sound examples from the italian point of view.

8.1.1 Methodology

In this section, we describe the process involved to analyse the papers in the sustainable manufacturing field. To give a more detailed vision of the topic, we consider this field divided in 6 sub-classes (one for each principle), adopting the well know 6R framework. The goal of the process is both to identify which the topics in each of the 6R are, then to accurately measure the way this framework corresponds to the topics of sustainable manufacturing. As seen in the flowchart of figure 1 the main activities of the process are four: Paper manual classification, Automatic Keywords Extraction, Keywords manual selection and Clustering. Each activity will be described in the next subsection.

Assignment of 6R principle meanings

The first critical issue is the selection criteria of 6R principles for the assignment. This was done by trying to assign a semantic specification of each R principle, as below explicated, built by summarizing all the possible definitions and concepts extracted by the selected bibliography (partially cited in the present paper). The semantic specification provided below refers to the functional scope of each class, intended as an activity to be performed.

Reuse

Reuse is the action or practice of using something again, whether for its original purpose (conventional reuse) or to fulfill a different function (creative reuse or repurposing). Reuse involves taking, but not reprocessing,

[1] http://www.trade.gov/green/sm-101-module.asp

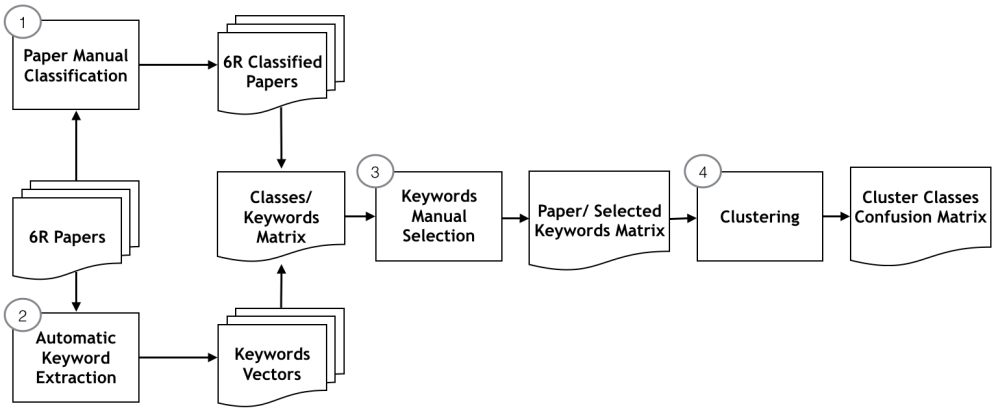

Figure 8.1: Flowchart of the 6R papers analysis process. The white boxes represent the activities and the yellow boxes the predicted documents. In the figure are also highlighted the input and the main outputs of the process.

previously used items in order to save time, money, energy, and resources. In particular reuse is useful for machineries, especially when they are expensive, and for their components. Besides, technologies can be also reused and readapted to different situations. In addition, reuse is important because it minimizes disposal needs and costs.

Reduce

In pre-manufacturing it is possible to implement the reduction minimizing the use of resources. During the production phase, it refers to the use of energy and material, for example, it may involve the use of lower cost materials, the elimination of unnecessary product characteristics, the reduction of overhead costs or the readjustment of product processes. It implies a decrease of costs. As a consequence, the derived effect of this principle is to minimize and optimize performance in terms of cost, time and waste and it focuses primarily on the first three stages of the four life cycle of the product before recalled.

Recycle

Recycle involves the process of converting material that would otherwise be considered waste, into new products and it corresponds to the breaking down of used items to make raw materials for the manufacture of new goods. Recycling can prevent the waste of potentially useful materials and reduce the consumption of fresh raw materials, thereby decreasing: energy usage, air pollution (from incineration) and water pollution (from landfilling). Recyclable materials include many kinds of glass, paper and cardboard, metal, plastic, tires, textiles and electronics.

Recover

Product recovery operations refer to operational processes (e.g. disassembling, sorting and cleaning) on products at the end of their use, in order to make them usable in n subsequent life-cycles. Differently from recycling, the product life-cycle is shortened, skipping all that phases of retreatment of wastes up to their second use. The risks of implementing product recovery operations is the uncertainty of returned product quality and quantity and for this reason recovering activities have only recently been considered in manufacture.

Redesign

The redesign activity involves the act of redesigning the next generation of products, which would use components, materials and resources recovered from the previous life-cycle or previous generation of products. It refers to the evaluation of ideas turning them into concrete innovative products obtained from products in their post-use phase.

Remanufacture

Remanufacture involves the re-processing of used products, to restore them to their original like-new state through the reuse of as many parts as possible without loss of functionality. It includes and exceeds the activity of recovering, because a remanufactured product should match the same customer expectation as a new one.

Despite this initial classification, some problem occurred for experts in assigning the papers to the different classes, due to the absence of the formalization that is needed to define unique criteria of belonging. This fact brought to the unbalance of the number of words within each class (corresponding to each principle), that drove to the fuzziness of the training set and, as a consequence, the difficulty in performing a real classification of scientific papers. This is the first critical issue of the followed approach so far, that may be overcome by alignment sessions of the decision makers.

Automatic Text analysis and Keywords extraction

In order to analyze the information coming from scientific papers belonging to topics encompassed within the 6R framework, we exploit text mining techniques, which employ methods from different fields of data mining to extract meaningful information 8. The second activity in our analysis process is thus devoted to transforming the set of papers in a set of numeric vectors to be elaborated by the clustering algorithm. To this aim, some text mining techniques are applied in sequence 12 to automatically extract meaningful information and knowledge from unstructured texts. The text mining process performed is summarized in the following. First, the information content of the document is converted into a structured form (vector space representation). In fact, most of text mining techniques are based on the idea that a document can be faithfully represented by the set of words contained in it (bag-of-words representation 9). According to this representation, each document j of a collection of documents is represented as an M-dimensional vector, where M is the number of words defined in the document collection, and w(tji) specifies the weight of the word ti in document j. The simplest weighting method assigns a binary value to w(tji), thus indicating the absence or the presence of the word ti, while other methods assign a real value to w(tji).

In the following, the text mining steps performed are described (for details aobut the activities see section 5:

- *Tokenization* is the first step of our text mining process, and consists in transforming a string of characters into a string of processing units called tokens that could be syllables, words, or phrases. Typically, with tokenization other operations are performed with the aim of making the text cleaner. Such operations are the removal of punctuation and other non-text characters, and the normalization of symbols (e.g., accents, apostrophes, hyphens, tabs and spaces). In the proposed system, the tokenizer removes all punctuation marks and splits each text into tokens corresponding to words, bigrams and trigrams.
- *Stop-word filtering* consists in eliminating the words which provide little or no information to the text analysis, or that, in our case, could make the clustering process fuzzier. Common stop-words belong to certain part of speech classes, such as articles, conjunctions, prepositions, pronouns, etc. Other stop-words are those that typically appear very often in sentences of the considered language (language-specific stop-words), or in the set of texts being analysed (domain-specific stop-words). In our case, this second group consists of typical words of the scientific articles such as “paper”, “state-of-the-art”, “present work”.
- *Stemming* is the process of reducing each word (i.e., token) to its root form. The purpose of this step is to group words with the same theme having closely related semantics. In the proposed system, the stemmer exploits the Snowball Stemmer for the English language, based on the Porter’s algorithm.
- *Feature representation* consists in building, for each text, the corresponding vector of numeric features. Indeed, in order to cluster the texts, we have to represent them in the same feature space. In particular, we consider the F- dimensional set of features corresponding to the set of relevant stems.

Manual keywords selection

Keyword selection procedure was performed by hand from a team of 4 experts in the scientific field of sustainable manufacturing and with consolidated expertise on research and innovation. In order to prevent misunderstandings and difference in the judgement, a preliminary analysis of the single attitudes to classification was statistically performed and misalignment were eliminated by a guided training, based on the contemporary evaluation of a same sample from different experts so as to align the judgement. This process was done, as above mentioned, based on the 6R explicit framework and sharing the meaning of each principle.

It is clear that assignment of keywords to classes is not a trivial task and susceptible of interpretation. Disambiguation process would be required to explain the criteria of affinity to a given class, hopefully by referring to the functional scope of each class. Still an ambiguity may result, which can hardly be removed without a more profound specification with appropriate examples, which is out of the scope of the present paper.

According to this procedure the 6R refer to the following keywords:

- *Recover*: recovery, planning processes, renewable resources, new approach, waste recovery; process scrap recovery; energy recovery; heat recovery; collection; separation; design for environment; material recover optimization; recovery logistics;
- *Recycle*: recycle, adhesive technologies, materials conversions, government regulation; (Advanced) recycling technologies; Recyclability; End of first life; Down cycle; New/Old process scraps; Material scraps; Landfill taxes; Recycling benefit awarding; Secondary material production; Embodied energy saving; Environmental legislation
- *Redesign*: machining, cutting, lubrication, environment technologies, cryogenic; Material-efficient design; Design for Environment; New Materials; Eco-friendly design; Eco-friendly materials; Material reduction; Light- weighting; Design for Disassembly
- *Reduce*: reduce, optimize, waste minimization, consumption, energy payback, pollution, reduction, emission, saving; Energy reduction; Waste reduction; Resource reduction; Material usage reduction; Process sustainability optimization; Energy efficiency; Heat efficiency; Manufacturing efficiency; Doing with less;
- *Remanufacturing*: remanufacturing, material processing, eco-efficiency, renewable, innovations, nanocomposites; Product renewal; Product upgrade; Reconditioning; Part replacement; Modularity; Disassembly; Inspection; Separation; Re-Assembly; Design for Remanufacturing
- *Reuse*: reuse, replacing, material flow analysis, eco-efficiency; Component reuse; product reconditioning; product upgrade; non-destructive recycling; reuse supply chain; product maintenance; product repair; product monitoring;

It is clear that the keyword extraction is not a trivial task, since the outcome of the process may range from a useless single word to a meaningful group of combined words that, on the other hand, could become too specific to be significant for the training set of the search engines.

Paper Clustering

Document clustering is the application of the general process of cluster analysis to texts. The practical applications of document clustering systems are several and vary from automatic document organization to automatic topic extraction. For further details see section 5.5.4. Independently from the kind of algorithm, before the clustering phase, each document has to be represented as a set of features. These features are typically the n-grams contained in documents, so a critical activity for clustering effectiveness is the n-grams extraction (or feature selection). This goal is achieved with a series of sub-activity. These are typically tokenization (the process of parsing text data into smaller units), stemming and lemmatization (reducing all tokens to its semantic base), removing stop-words (less important words), and finally computing term frequencies (or other measures of the relationship between documents and words).

To get an exploratory view of the degree of precision with which the 6R framework represents the papers in analysis, the manually classified documents were clustered using a clustering Spherical K-Means Cluster-

ing algorithm (Buchta et al., 2012). We will then compare the output of the algorithm with the manual classification in the following paragraph.

Spherical k-means exploit cosine dissimilarities to perform prototype-based partitioning of term weight representations of the documents. Prototypes are centroids defined in the same feature space of the documents. The aim of the algorithm is to minimize, given a set of objects (documents in our case) and prototypes described in the same features space (the selected keywords) the cosine distance between each element and the closest prototype. This implies that the output of the algorithm is a membership matrix, in which each document is assigned to a certain prototype.

8.1.2 Results

In the present section, we describe the output of the application of the methodology described in section **??** on the 339 selected scientific papers on sustainable manufacturing extracted by the query "sustainable manufacturing" in the paper search field of the SCOPUS® database. Accordingly, apart the deliberate selection of the source database, no other filter was applied in the journal selection and this guarantees the significance of the journal sample selected. The main output, as highlighted in Figure 1, is the distribution of documents among the 6R principles (i.e. classes) as manually classified, the automatic keywords representations of the classes and the outputs of the clustering algorithm (which is an unsupervised assignment of the documents) compared to the manual classification.

The assignment of a paper to a class was made by recognizing the application, or tools or scope of the paper to one 6R's principles. In this sense, for instance, a paper belonging to a recover might have as scope, or keywords or approach or an activity performed with the principal task to assure the recovery of a give good. It is thus clear, from the beginning, how complex may be to recognize just one principle more than a multiple possibility to satisfy more principles. This fact will be discussed in the followings.

Distributions of documents among the 6R

The main point coming out from the histogram in Figure 8.2 is that the distribution of scientific papers assigned to the 6R principles (classes) is not homogeneous. If the four classes Reduce, Recycle, Redesign and Remanufacture collect a comparable number of papers, the other two (Reuse, Recover) are poorly represented. One possible explanation of this is that cases of reuses or recoveries are less generalizable compared to other "R", and thus potentially less interesting for conferences and journals. Recovery and reuse, in fact, often refers to specific functions embedded into the product, and finding a general rule for assessing or designing such processes might be more difficult. They rather might find room in technical magazines, not included in the present analysis as concern processes. It worth noting that reuse and recover here refers to product and not explicitly to materials. Different approach concerns the design for reuse, which to a certain extent would belong to remanufacturing, as in the most of papers discussed, to mention a few, in the Annals of CIRP 18,19,20. It is difficult to think that the classification criteria, which plays a critical role into the assignment to classes, contributed to this situation, depending on the meaning assigned to each principle. Whether a different combination of classed have been provided, a different histogram would be expected to appear.

Keyword vector representation of 6R

After the phases of automatic keyword extraction and manual keyword selection, we derived the paper/selected keywords matrix. A sample of the elements of this matrix (keyword and their occurrency) is:

- *Recover*: Life cycle 5, environmental impact 4, raw material 3, cycle assessement 3, tossii fuel 3, energy payback 2, cleaner products 2, energy cost 2, energy usage 2, pv System 2, product process 2, managements System 2, new technologies 2, energy save 2, climate change 2, risk assessements 2
- *Recycle*:lite cycle 40, environmental impact 31, waste management 30, resource conservation recycle 25, solid waste 21, lite cycle assessement 18, recycle material 14, recycle process 12, reuse recycle 11,

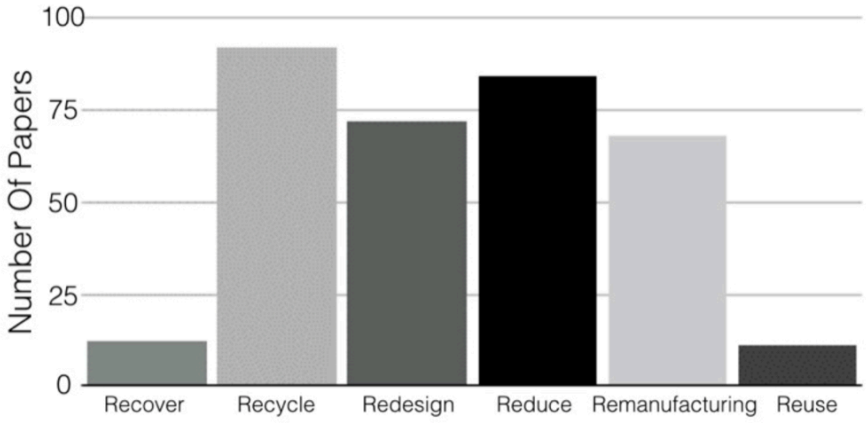

Figure 8.2: Histogram of the number of papers manually assigned to each of the 6R principles.

electronic equipement 11, global warm 9, waste disposai 9, heavy metal 8, environment friendly 8, waste treatment 8, waste stream 8

- *Redesign*: manufacturing process 20, lite cycle assess 19, sustainable manufacturing 17, machine tool 14, naturai gas 14, cutting tool 13, raw material 13, tool life 13, cutting fluid 13, tool wear 13, mechanical engineering 12, System boundaries 11, machining process 11, manufacturing System 11, cutting speed 11, cutting zone 10
- *Reduce*: life cycle 40, environment impact 38, energy consumption 31, life cycle assessment 28, energy usage 22, energy requirements 18, solar energy 15, cleaner product 13, global warm 13, climate changes 12, solid waste 12, gas emission 11, solar celi 11, cutting conditions 11, co2 emissions 10, environmental management 10
- *Remanufacturing*: sustainable manufacturing 28, manufacturing process 22, remanufacture product 18, supply chain 16, new product development 16, climate change 12, business model 12, product remanufacturing 12, manufacturing System 11, environmental performances 10, remanufacturing industry 10, management System 9, waste generation 9, design for environment 9, remanufacture and recycle 8, automotive industry 7
- *Reuse*: life cycle 7, environment impact 7, energy usage 4, mechanical engineering 4, industriai ecology 4, cleaner product 3, life cycle assessement 3, environment management 3, design for manufacturing 3, environment protection agency 3, waste management 3, united nation 3, environment management System 2, process design 2, sustainable manufacturing 2, green chemistry 2

The keywords were divided according to the 6R principles (i.e. class), listing the number of the paper belonging to each class.The sample list of keywords and their occurrency, reflects the distribution of figure 8.2 and cannot thus be significant for classification but only for clustering purposes at the present.

Clusters/classes confusion matrix

To analyze from a semantic dimension how well 6R framework represent sustainable manufacturing definition through the scientific papers considered, the manually classified documents were clustered using a clustering Spherical K-Means Clustering algorithm. The results (the assignments of each paper to a cluster) was then compared with the manual classification. The number of resulted clusters was 6, having this number of clusters no specific relation with the number of 6R classes. The output of the clustering process is thus a 6x6 Matrix (Class(row)/Cluster (column)) shown in Figure 8.3. Numbers in the cells of the matrix represents centroids that can be characterized by the related keywords.

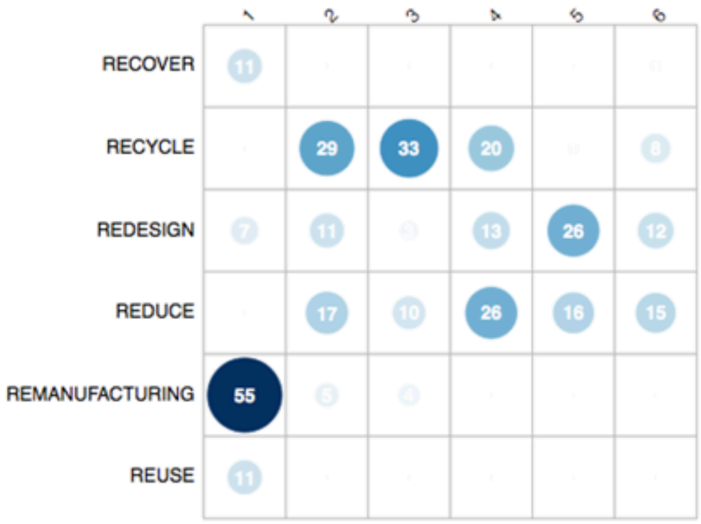

Figure 8.3: Matrix: comparing the manual classification and the clustering output.

8.2 Sustainable Manufacturing: An Extended Mapping

The methodology described in section **??** has been then re-adapted to give an extended anlysis of the 6R framework from a bottom-up perspective. In the present section, we map the state of the art of Sustainable Manufacturing using topic modelling, a widely used text mining techniques. The aim is to give a broad overview of how this knowledge field is approached world-wide and then, in section 3, to have a deeper look at the Italian way to sustainable manufacturing.

8.2.1 Methodology

The process of topic extraction is composed by the sequent activities:

1. *Papers Collection*
2. *Keyword Extraction*
3. *Topic Modelling:* Number of topic selection
4. *Topic Modelling:* LDA Model fitting
5. *Manual topic labelling*

Each activity is described in the sequent sections.

Papers collection

The analysis starts form a corpus of papers on sustainable manufacturing. The papers were downloaded from the Scopus database searching for the query:

$$TITLE - ABS - KEY(sustainable. * manufacturing)$$

At the date of 29/05/2018 the query results in 1,628 documents.

Keyword extraction

We represent each article as a set of keywords, merging Author Keywords and Index Keywords. The keywords are then "sanitized" following the sequent rules:

- Eliminate duplicated keywords
- Eliminate brackets and its content
- Substitute non-alphanumeric character with a blank space
- Merge synonyms, alternative spelling and keywords pointing to similar concepts
- Eliminate scientific literature specific keywords like "article" and "review"
- Filter the generic keywords. The metrics for the threshold is the percentage of papers that contains the keywords. The value has been set to 7.5%.

Finally, we compute the Keywords Term Matrix. The matrix is composed of 1,628 documents and 26,272 keywords. Having a mathematical representation of the documents allows us to apply standard mathematical techniques to them.

Number of Topic Selection

The goal of the present section is to compute a topic model based on the keyword representation of the papers. A topic model allows examining a set of documents and discovering, based on the statistics of the words in each, what the topics might be and what each document's balance of topics is. To deploy it for the purposes of the present paper, keywords will cluster to form a topic: each topic will represent a different view (say, principle) of sustainable manufacturing. This approach brings a new perspective to the definition of sustainable manufacturing with respect to the 6R's framework.

To compute the topic model, we use the Latent Dirichlet Allocation (LDA) algorithm. To select the number of topics for LDA, the most efficient and effective way is to calculate multiple metrics of the quality of the results in function of the number of topics. All existing methods require to train multiple LDA models to select one with the best performance. Several approaches tried to take the problem of automatically finding the right number of topics contained in a set of documents. Every approach follows the idea of computing distances (or similarities) between pair of topics varying the number of topics. We used four different methods to evaluate the output of a topic model for different value of k. These methods are:

- Caojuan2009 (Cao et al., 2009): Minimize the average cosine distance between every pair of topics. The best topic number K has a minimal final distance between topics in Latent Dirichlet Allocation
- Arun2010 (Arun et al., 2010): Minimize the symmetric KL-Divergence of the salient distribution that are derived from the matrices of factors. These matrices are the re-projections of the documents on the topics and of the topics on the vocabulary (the selected tokens). The divergence values are higher for non-optimal K values.
- Griffiths2004 (Griffiths and Steyvers, 2004): Maximize the likely of the data given the model built considering K topics. This is a problem of model selection using Bayesian statistics.

We computed these for measures fitting a LDA model for every K between 2 and 20. We choose the higher value because we wanted to obtain a reasonable number of topics representing the concepts (or we may say, principles) behind sustainable manufacturing. The results of the analysis are shown in figure 8.4. From this figure is evident how the measure we want to minimize intersect each other at a value between 11 and 12 topics. We thus decided to visualize the results of the LDA models for a number of topics k between 7 and 15: the expert panel interviewed then allowed to decide that the best results was at 12 topics.

8.2.2 Results

We then extract the one-topic-per-term-per-row probabilities, beta. Beta measures what is the probability for each topic to produce a term. Figure 8.5 shows the top 5 terms for each topic. Here each topic has a different vertical position and a different color. Each topic is represented by a set of keywords and each keyword has a different position on the y axes. For each intersection topic/keyword, the dimension of a circle

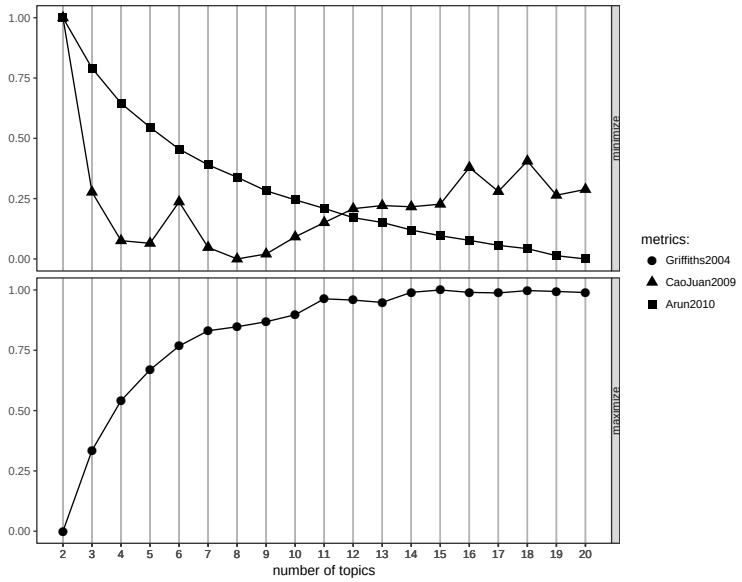

Figure 8.4: Relationship between the measures of the quality of the topic modelling output and the number of topics.

represent beta (what is the probability that topic to produce that term). It is possible for some keywords to belong to multiple topics: for example, optimization belongs to topic 1 and 9. It is easy to spot these keywords because typically (belonging to two different topics) are far from the group of keywords of the topics they belong.

Manual Labelling of the topics

To have a clearer representation of the topics, these were manually labelled by 6 independent experts in the field of manufacturing and sustainability. Then a seventh experts took in input the results of the manual naming and synthetized the label. The final results are the names of the 12 topics. These names are:

1. Smartness for Sustainability
2. Sustainable Machining
3. Manufacturing Environmental Efficiency
4. Modelling Manufacturing Sustainability
5. Welding Sustainability
6. AM Sustainability
7. Life-cycle Product Management
8. Advanced Material Sustainability
9. Production Management for Sustainability
10. Sustainable Energies
11. Innovation for Sustainability
12. Sustainable Logistic

This final picture provides a clear representation of how topics recognized are mostly homogeneous, provided there are few outlier keywords per each topic, as in figure 8.5. This “new perspective” of sustainable manufacturing represents a way of defining its contents by the grace of the tools or domain of interest involved. For instance, topic n.1 (smart sustainability) refers to the use of the new group of tools and approaches belonging to the Industry 4.0 stream, thus defining specific contents and criteria useful to pursue the sustainability in manufacturing.

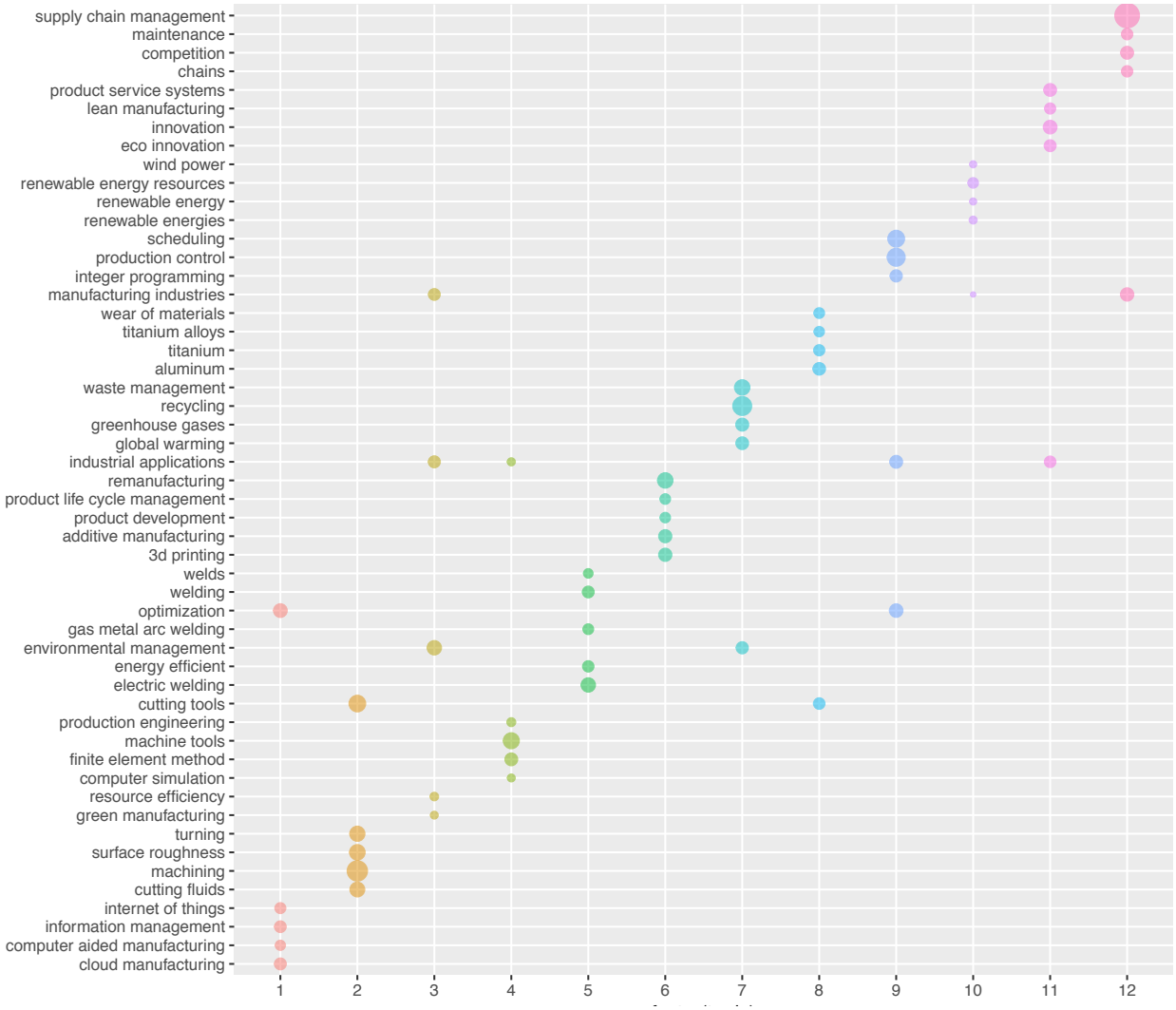

Figure 8.5: Content of the 12 topics. Each topic has a different vertical position and a different color. Each topic is represented by a set of keywords and each keyword has a different position on the y axes. For each intersection topic/keyword, the dimension of a circle represent beta (what is the probability that topic to produce that term).

8.3 Blockchain

In section 7.2 has been proposed a new usage of sentiment analysis to extract advantages (considered synonym of the words benefit, gain or profit) and disadvantages (drawback, failure) of technologies from patents with the aim of helping researchers and designers to effectively develop new products, analyzing positive and negative properties of an innovation. The main contributions of the present analysis is to list the problems of a technological field using this sentiment tool. We want to understand which are the problems that research is trying to solve for a technology and if is there any focus from researchers on the solution of certain technical problem. Furthermore using text mining techniques it is possibile to correlate these problems between them and understand if is there any correlation between problems of technologies and if this map help researchers and companies to understand the research agenda.

This section thus proposes an innovative methodology for the (semi-automatic) extraction of technical knowledge from academic articles, using text mining techniques. Among many information related to technical knowledge, thw methodology focus on the collection and the analysis of the problems of a technology. The meaning of problems of a technology is related to the concept of disadvantages of a technology but with a broader sense. In fact, using papers as a source makes it possible to collect not only technical disadvantages (typically described in patents) but also, organizational, business or even social problems, since these are faced by different intellectual fields in the scientific literature. We also propose a case study on the application of the method to the field of *block-chain*. We chose this technology because:

- It is highly innovative
- It is having an impact in many different field
- The problems of this technology are both technical and managerial

How Blockchain Technology Works

The blockchain can be exemplified as a process in which a set of subjects shares computer resources (memory, CPU, band) to make available all private users in which each participant has a copy of the data.

The use of cryptographic validation techniques generates the mutual trust of the participants in the data stored by the blockchain, which makes it comparable to the registries managed in a centralized manner by recognized and regulated authorities (banks, insurance companies, etc.) (Pilkington, 2016).

A blockchain is an open and distributed register that can store transactions between parties in a safe, verifiable and permanent way. Once written, the data in a block can not be retroactively altered without the modification of all the subsequent blocks, and this, due to the nature of the protocol and the validation scheme, would require the consensus of the majority of the network. (Iansiti and Lakhani, 2017)

The blockchain is a continuously growing list of records, called blocks, which are linked to each other and made safe by the use of cryptography. Each block in the chain contains a hash pointer (that is a link to the previous block), a timestamp and the transaction data.

The distributed nature and the cooperative model makes the validation process robust and secure, but it has considerable time and costs, due in large part to the *price of the electricity* needed to validate the blocks (Underwood, 2016). Authentication takes place through mass collaboration and is fueled by collective interests. The result of all this is a robust workflow where participants' data security expertise is not required. The use of this technology also makes it possible to overcome the problem of the infinite reproducibility of a digital asset and of the double expense without using a central server or an authority. (Karame et al., 2012)

There exist an overtrust about blochain technology for businesess, and there exist a growing interest about the drawbacks of this technology (Eyal and Sirer, 2018; Lin and Liao, 2017; Yli-Huumo et al., 2016). Has been pointed out that a 51 percent attack 26 would be enough to access or control a private blockchain because, in most cases, the organization that controls it already controls one hundred percent of the block creation. In fact, if someone could attack or damage the blocking tool on a private company server, would be possibile to get complete management of the network and the ability to access and modify the data (Hampton, 2016). This is because the centralization due to the privatization of the blockchain leads to a

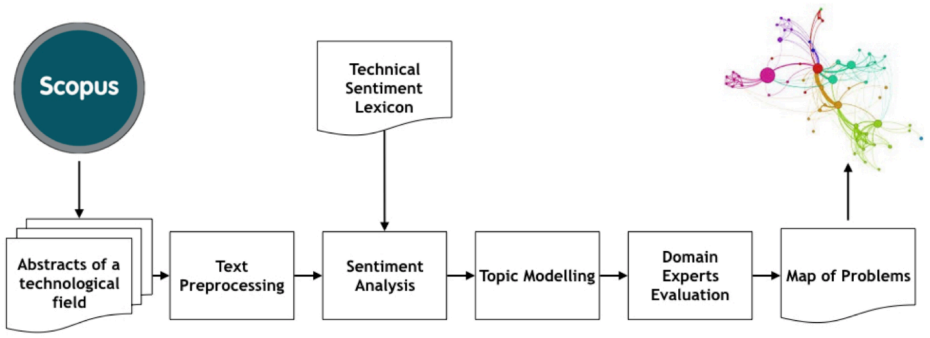

Figure 8.6: Proposed workflow for the problem extraction from papers.

single point of failure, or a central break point, which in a public blockchain could not happen as distributed, so there would be no central point to attack. This could have profound implications in financial crises or debt crises like the 2007-08 financial crisis.

8.3.1 Methodology

Considering our objectives, the approach we adopted for paper analysis is radically new with respect to the one traditionally established in the literature (bibliometric analysis and keyword approaches). These approaches do not allow a deeper understanding of the technical content described in the text. For this reason, we relied on text mining techniques supported by our technical knowledge base. For the sentiment computation in fact, we used a technical sentiment lexicon developed by the authors that extracts advantages and disadvantages of inventions from patents (7.2) and a novel dictionary lookup approach that incorporates weighting for valence shifters. Following a bottom up-approach, we redesigned and updated these lexicons for an optimal application to paper documents.

The workflow of the proposed methodology is shown in figure 8.6. The process starts with the collection of abstracts belonging to the same technological field. The abstracts are downloaded using the Scopus API, extracting all the documents that contains keywords of the selected technology in the title, abstract or keywords fields. The texts are then pre-processed using state of the art natural language processing tools (sentence splitter, tokenizer, lemmatization). Then, for each sentence we computed a negative sentiment polarity and took into consideration only the sentences having a negative polarity score below a threshold level. We applied topic modelling algorithm on the negative sentences with the aim of clustering. The output of topic modelling is evaluated by technology domain experts with the aim of labeling the identified clusters.

The application scenarios of our methodology have a wide range of users:

1. Companies that want to rapidly map a certain technological field
2. Policy maker that want to invest to solve problems of a technology to boost its innovation
3. Journals, to understand the hot-topic of a specific field
4. Research networks, to exploit possible collaborations and synergies between scholars.

Document collection and selection

The documents were extracted from Scopus searching for the query:

$$TITLE - ABS - KEY(blockchainORblockchainORblock - chain)$$

At the date of 29/05/2018 the query results in 1,628 documents.

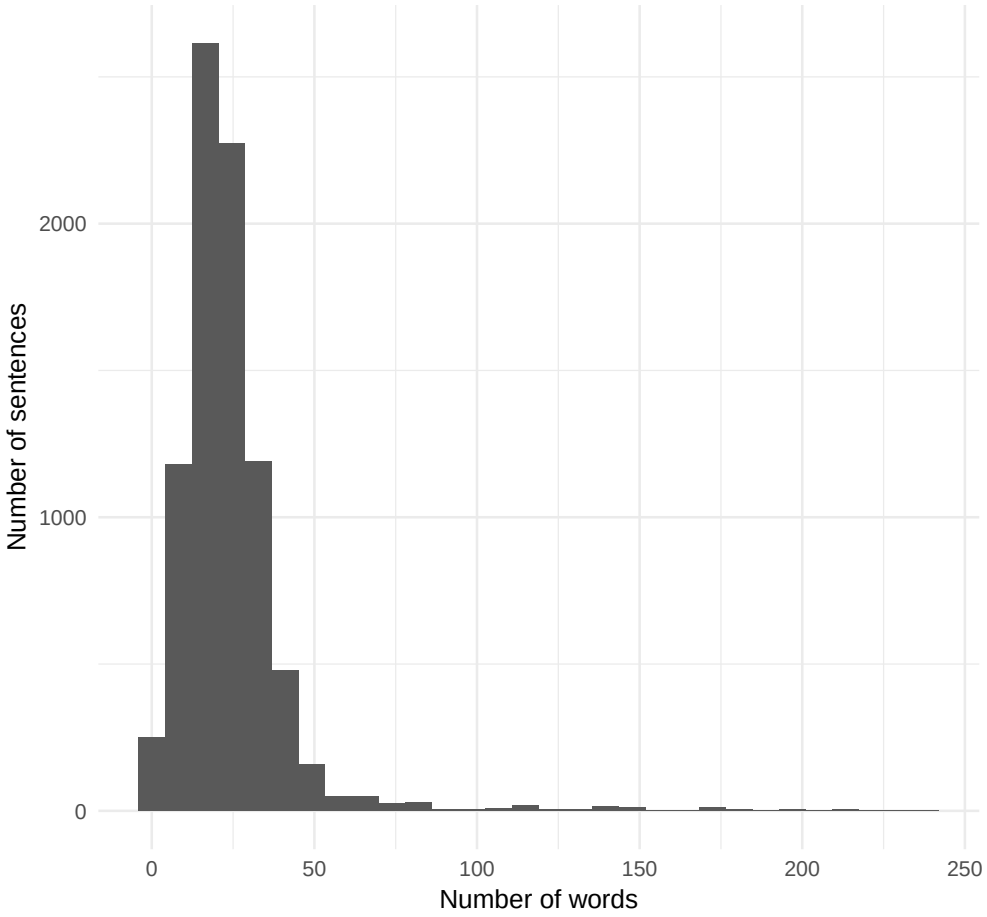

Figure 8.7: Histogram of the lengths of the sentences.

The term blockchain and its morphological variations is used in other fields different from computer science and economics, in particular in chemistry. For this reason we filtered from the data set all the papers belonging to one of the sequent All Science Journal Classification (ASJC) classes: chemistry, physics and astronomy, chemical engineering, biochemistry, genetics and molecular biology, pharmacology, toxicology and pharmaceutics, immunology and microbiology. The result of this lead us to a set of 1,364 papers.

To have a dataset with an higher precision, we also filtered using rules based on the content of the abstracts, since many conferences do not have an ASJC class in scopus. This characteristic of the scopus database could lead to bias in the results, especially for innovative technologies like block-chain for which the scientific discussion is stronger in conferences then in more structured journals. For this reason we analyzed the abstracts of the 264 articles belonging to one of the ASJC classes listed above (chemistry related) extracting the most frequent words and filtering for a blacklist of generic words contained in scientific articles. This results in a dictionary of words that has ben used to search and filter al the papers containing in the abstract one or more of the words in the dictionary. The final number of selected papers is 1,276.

Sentence Splitting

A sentence splitter (for furhter details see setction 5.4.1) is then applied to the abstract to divide the texts in sentences. From this step we had an output of 8,406 sentences.

The next step involved the measure of the sentences length. Some of the sentences in fact could be too long due to several fact for example the style of the author or an error in the sentences splitting phase). Since it

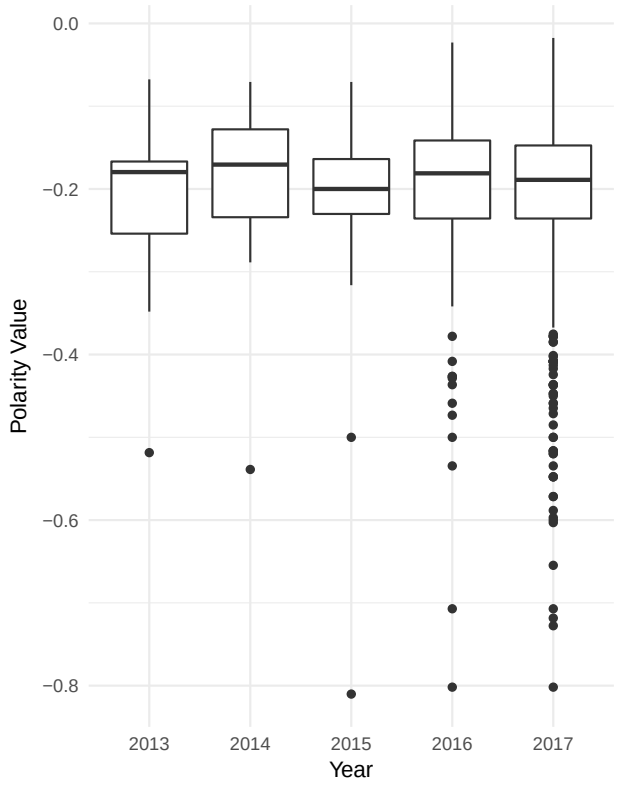

Figure 8.8: Box plots of the distributions by years of the polarities of the sentences having a negative polarity score.

is not a goal of the present work to analyze the style of the authors or to design a better sentence splitter, we decided to filter the sentences that are too long. The decision on the threshold level of number of word has been taken analyzing the distribution shown in figure 8.7. The 97% of the population of sentences contain a number of words that is lower than 53. Thus all the sentences contain more than 53 words has been filtered. From this step we had an output of 8143 sentences.

Polarity Computation

At this point has been possibile to apply a sentiment polarity measurement on the sentences (Rinker, 2018). As output we had a polarity score between -1 (strongly negative) and 1 (strongly positive) for each sentence. Since we are interested in extracting the problems of the state-of-the-art blockchain systems, we visualize the distribution of the polarity scores of the sentences labeled as negative (having a polarity score lower then 0). The histogram of this distribution is shown in figure TOT. In this histogram are represented 1,108 different sentences. As we can see we have a distribution center in -0.2, with a tail that goes down to -0.8. This is an evidence that the system give a reasonable measures since it would be unreasonable to have more probable extreme values. Furthermore, the number of sentences having a polarity equal to 0 is 4,435 that is about the 50% of the sentences and the number of sentences having a positive polarity is 2,700 : it is reasonable that the higher number of sentences falls in these classes since the scientific jargon has to be neutral. The positive bias has been seen also for patent documents (see section 7.2)).

One interesting evidence we found is that the number of outliers (having a polarity values lower than the 95% of the population) is growing over the year as shown in figure 8.8. Even if the mean remains near to -0.2, int he last five years the number of strongly negative sentences has increased. This shows how there has been a trend in the focus on the problems of blockchain.

Tokenization and Token Filtering

In the next phase we apply a state of the art natural language processing pipeline (see section 5 for more details) to the 1,108 negative sentences.

The tokenization process results in 220,094 tokens. To extract the meaningful words (words that adds information about the problem that the sentence is describing) we applied a series of filtering steps. As meaningful unit to analyze (token) we decided to use the lemma of the word. The 9,451 final token is a reasonable number considering the 1,108 sentences in input: we have a mean of 9 words per sentence. Considering the mean of 25 words per sentence shown in figure 8.7, we now have a summary of the sentences contains only the meaningful words. This output is a clean input for the topic modelling phase.

Topic modelling

Topic modeling is a method for unsupervised classification of documents which finds groups of documents fitting a statistical model to the data. In other words, these models captures word correlations in a collection of documents with a set of topics. Latent Dirichlet allocation (LDA) is a particularly popular method for fitting a topic model. It treats each document as a mixture of topics, and each topic as a mixture of words. This allows documents to "overlap" each other in terms of content, rather than being separated into discrete groups, in a way that mirrors typical use of natural language (for further details see section 5.5.6). If the topic modelling process is good, we have as output a structure where every topic is an understandable, meaningful and compact semantic cluster that in our case clearly represent a state-of-the-art problem of block chain technology.

Document Term Matrix

The first step is to take in input the 9,451 tokens and count how many times each one occurs in the 1,108 sentences. In this way we obtain a document-term matrix (DTM) where the documents are the sentences and the terms are our tokens (the lemmas). In our case the DTM has 1077 rows representing the sentences (31 sentences did not contain any relevant tokens) and 1147 tokens.

Definition of the Optimal Number of Topics: a Human-Machine Hybrid Approach

To fit a LDA model we have to give as a parameter the number of topic k that we think that best represent the corpus in analysis. In literature there exists many measurement to compute an optimal value of k. However it is worth to keep in mind that these measures are not always correlated with expert judgement about topics interpretability and coherence. For this reason in the present paper we use an hybrid approach in which we find an optimal neighborhood of value using state of the art k tuning methods and then we compute a graphical output for the 5 best k values to be evaluated by 4 different experts.

Several approaches tried to take the problem of automatically finding the right number of topics contained in a set of documents. Every approach follow the idea of computing distances (or similarities) between pair of topics varying the number of topics. We used four different methods to evaluate the output of a topic model for different value of k. These methods are:

- *Caojuan2009* (Cao et al., 2009): Minimize the average cosine distance between every pair of topics. The best topic number K has a minimal final distance between topics in Latent Dirichlet Allocation
- *Arun2010* (Arun et al., 2010): Minimize the symmetric KL-Divergence of the salient distribution that are derived from the matrices of factors. These matrices are the re-projections of the documents on the topics and of the topics on the vocabulary (the selected tokens). The divergence values are higher for non-optimal K values.
- *Griffiths2004* (Griffiths and Steyvers, 2004): Maximize the likely of the data given the model built considering K topics. This is a problem of model selection using Bayesian statistics.

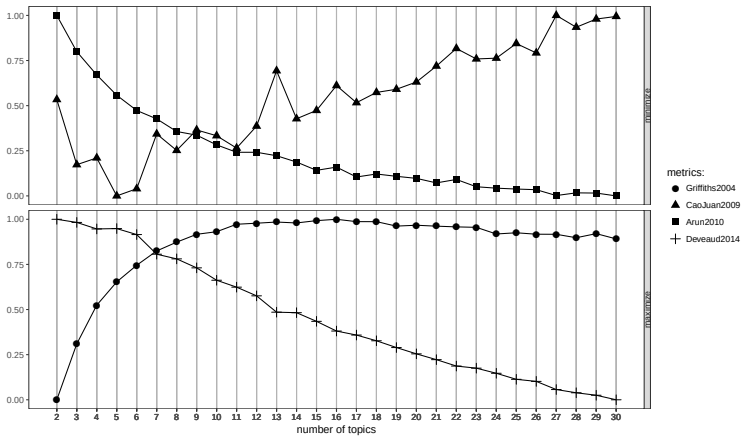

Figure 8.9: Measures of quality of the topic modelling results for growing number of topics. The figure is divided in metrics to minimize (top figure) and to maximized (bottom figure).

- *Deveaud2014* (Deveaud et al., 2014): Maximize the information divergence between all pairs of topics. The optimal value k is the value for which LDA modeled the most scattered topics.

We computed these for measures fitting a LDA model for every K between 2 and 30. We choose the higher value because we wanted to obtain a reasonable number of topics representing the problems of blockchain. After a brain storming with experts it emerges that there can not exists more than 15 classes of problems. We decide to take the double of this number to take in to consideration the biases of human experts. The results of the analysis are shown in figure 8.9. From this figure is evident how the measure we want to minimize interest each other at a value of 7 topics; the values we want to minimize are a little more unstable (especially Arun2010) and intersect both at level of 9 and 11. We thus decided to visualize and make choose the results of the LDA models for a number of topics k between 7 an 10.

To help experts in their decision process, an interactive map of the problem hass been generated for 7, 8, 9, 10 and 11 topics. The map was created using Shiny (Chang et al., 2017). The interactive map of problems (a screen shot of the app is shown in figure 8.10) helped experts to understand the relationships between the problems. Here a multidimensional scaling algorithm compute the inter-topic distance that permits to visualize the distribution of topics in two dimensions. Clicking on a topic the experts can see its content, and for each word can see both at the estimated term frequency within the selected topic and at the overall term frequency.

The experts agreed for a total number of 8 topics as optimal.

8.3.2 Results

The output topic modelling phase is show in figure 8.11. The results of the topic modelling are visualized to make it possibile for the expert to explore and label the topic representing the problems of block chain technology. Each point is a word, and word belonging to the same topic are grouped. The size of the label is proportional to the probability ß that the word belong to that topic.

8.4 Precision Agriculture

The agriculture is facing rapidly changing economic, social and environmental scenarios in 21st Century. Forecasts on worldwide population estimate that it might increase at about 9 billion by 2050 and led the Food and Agriculture Organization of the United Nations (FAO) to underline the growth of food needs of about 60% if compared to the annual average calculated between 2005 and 2007. The Committee on

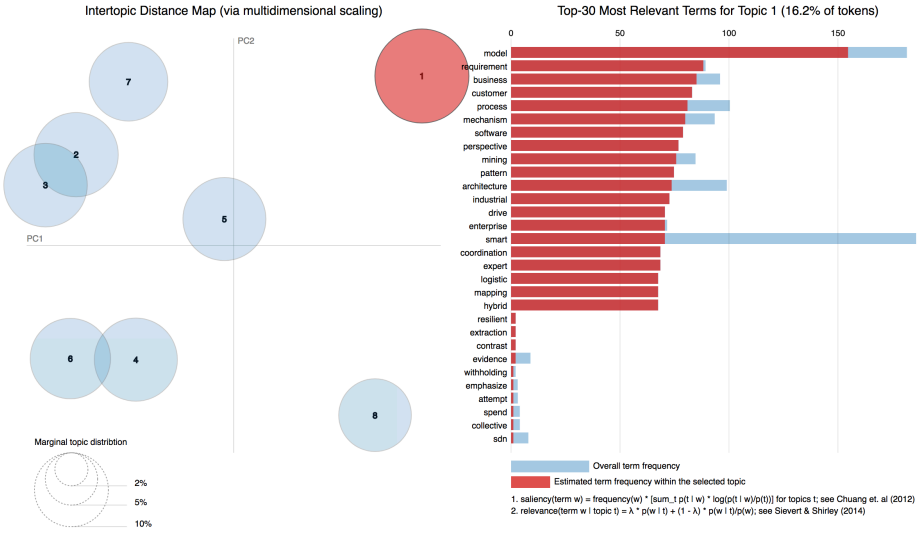

Figure 8.10: Screenshot of the shiny app used by experts to choose the optimal number of topcis.

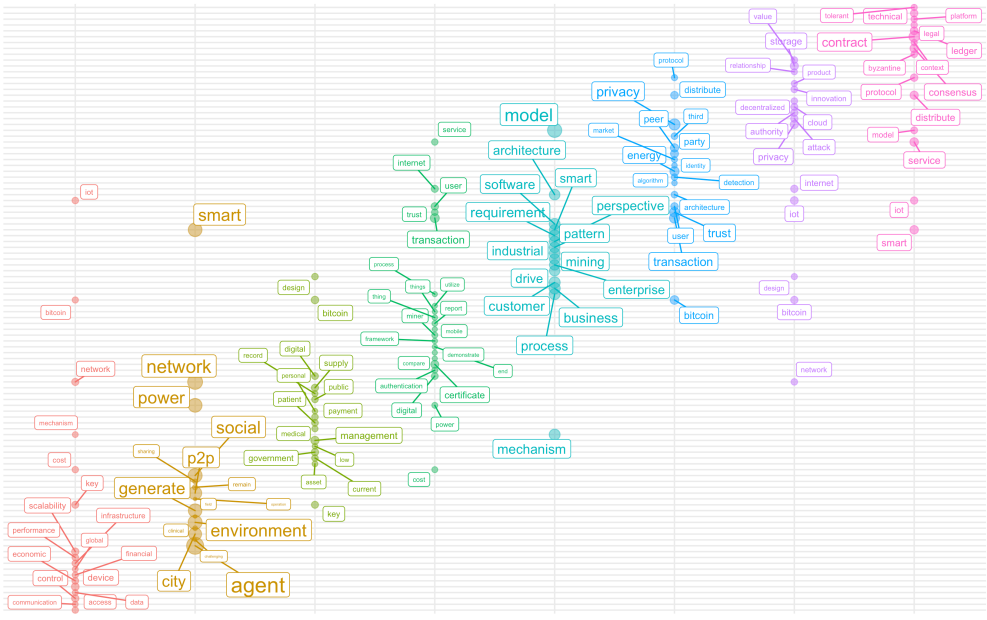

Figure 8.11: Content of the 8 topics of problems of blockchain. Each topic has a different vertical position and a different color. Each topic is represented by a set of keywords and each keyword has a different position on the y axes. For each intersection topic/keyword, the dimension of a circle represent beta (what is the probability that topic to produce that term

Agriculture and Rural Development of the European Parliament confirms these figures. In this regard, FAO focuses on how global agricultural and food systems can support the subsistence needs of the worldwide population according to different cultures and dietary habits of both developed and emerging countries.

Compared to past agricultural policies, nowadays trends related to technological evolution, socio-political changes, water shortages, as well as the increase in energy needs and the emergence of new pests and diseases that affect agricultural production acquire greater importance in the discussion between policy makers and companies concerning the future of agricultural activities.

In particular, technological innovation is necessary for the development of companies which aim to strengthen their production processes and organizational structures by exploiting automation and ICT within the cultivation and commercialization processes. Thus, digital technologies at the base of Precision Agriculture are the assets to leverage to deal with two major challenges for modern agriculture. On one hand, the need for an increase in production quantity by optimizing production factors. On the other hand, complying with production standards by combining appropriate quality levels and limited environmental impact. Precision Agriculture is a modern approach for agricultural management that exploits cutting-edge technologies to monitor and optimize agricultural production processes.

The concept of Precision Agriculture was born in the United States in the early nineties, where the House of Representatives (Liaghat et al., 2010) defines it as "an integrated information- and production-based farming system that is designed to increase long term, site-specific and whole farm production efficiency, productivity and profitability while minimizing unintended impacts on wildlife and the environment". Although it is a relatively well-known concept, Precision Agriculture still presents a low rate of adoption as reported by both academic surveys and professional reports (Pierpaoli et al., 2013).

Different definitions given by researchers, practitioners and policy makers have gradually allowed to deepen the understanding of the constituent elements of the concept as shown in the following table. Table 1 presents a literature review carried out on the Scopus database focusing on the analysis of the definitions of the concept of Precision Agriculture, and shows how technology emerges as the enabling aspect of Precision Agriculture. Anyway, authors focused also on other elements that refers to Precision Agriculture.

Pierce and Nowak (Pierce and Nowak, 1999) highlight the centrality of Technology and Benefits as the two key elements of Precision Agriculture. Zhang et al. (Zhang et al., 2002) strengthen these two distinctive elements, which compare in all subsequent analyses (Stafford, 2000; Kirchmann and Thorvaldsson, 2000), until the introduction of the concept of Sustainability (Bongiovanni and Lowenberg-DeBoer, 2004). Indeed, Bongiovanni emphasizes the environmental topic by underlining the role of Precision Agriculture to manage harvest production inputs in an environmentally friendly manner. Finally, Gertsis et al. (Gertsis and Vasilikiotis, 2018), define Precision Agriculture as "A modern farming management concept using digital techniques to monitor and optimize agricultural production processes". They introduce the concepts of digital techniques and optimization of production processes by highlighting some possible concrete applications of Precision Agriculture. As emerged from the previous analysis, a fundamental role for the implementation of Precision Agriculture is played by digital technologies. In particular, object identification, georeferencing, measurement of physical and chemical parameters, satellite navigation, connectivity, data storage and analysis, process automation and vehicle driving are the most adopted (Schrijver et al., 2016). Precision agriculture is then based on a cyclical process of observation and acquisition of data, followed by an interpretation and evaluation of the information acquired, and by the implementation of a series of decisions that respond to them. Thanks to these technologies, farmers can increase production, optimize resources consumption (workforce included), costs and quantitative and qualitative production possibilities, according to the specific characteristics of the soil and cultivation. (Sawanta et al., 2016) shows that the implementation of digital technology in agriculture can increase total profitability from $ 55 to $ 110 per acre (1 acre = 4046.87 m^2).

However, whether the Precision Agriculture is still far from being widely adopted (Pierpaoli et al., 2013; Tey and Brindal, 2012). Indeed, the implementation difficulties related to the high initial investments and the lack of suitable skills among the farmers identified still represent significant obstacles. In fact, a study conducted by Pierpaoli et al. (Pierpaoli et al., 2013) identifies the so-called "non-adopters" farmers, (e.g. those who do not have sufficient skills or financial resources to manage the PA's instruments). For these

reasons measuring the concrete benefits of Precision Agriculture is not an easy task. However, Precision Agriculture is an increasingly pervasive concept within the agricultural sector and the constant evolution of technologies linked to it generates many opportunities which have not been fully explored yet.

Technologies advancements are directly linked also to another pervasive innovative concept in worldwide economy: Industry 4.0. This new paradigm grounds on the exploitation of digital technologies in the development of business processes and its enabling technologies are also used in Precision Agriculture . For example, Cyber-Physical Systems can be seen as one of the cornerstones for the development of innovative solutions to monitor and manage processes in agricultural businesses.

Starting from this relationship, the present chapter aims at analysing the technologies at the bases of the two domains to identify possible overlaps between Precision Agriculture and Industry 4.0.

8.4.1 Methodology

The work focuses on the creation of a dictionary which identifies the most innovative technologies that are applied in Precision Agriculture by investigating the overlaps with Industry 4.0 technologies to create clusters and to analyze the connections between them.

The dictionary aims to analyze the technologies related to the Precision Agriculture domain and to identify those belonging also to the Industry 4.0 paradigm. Concretely, the dictionary is a list of Precision Agriculture technologies that were identified by analysing papers retrieved from the international database. First of all, the twenty-five most cited scientific papers were identified by using the query "precision agriculture", among those published between 2002 and 2017. Secondly, a text mining analysis on the 25 papers was conducted to identify the technologies mentioned and belonging to the Precision Agriculture domain. However, this list of technologies was not to considered exhaustive because of the reduced number of analyzed sources, which however provided the basic information to build the analysis context. To face this limitation, and given the proximity of the founding concepts of Industry 4.0 and Precision Agriculture, the Technimetro® was used to expand the list of technologies at the base of the dictionary. Technimetro® (Chiarello et al., 2018) is a dictionary that contains over 1500 technologies belonging to the Industry 4.0 paradigm and was created by selecting the Industry 4.0 technologies found in manuals, technical documents and scientific publications on Scopus. The relationships between these technologies were studied through a text mining activity to describe possible clusters and to understand how technologies are linked one-another. Therefore, the present work attempted to understand if the Technimetro® contains other Precision Agriculture technologies that were not identified with the analysis of the papers.

To answer this question, all abstracts of the publications on Precision Agriculture (published on SCOPUS from 2002 to 2017 for a total number of 4320 papers) were analyzed using the software "R". In this way, technologies belonging to the Technimetro® that were mentioned in the Precision Agriculture publications were identified. Therefore, technologies extracted through the Technimetro® have been checked to manually eliminate not applicable terms. Technologies ware removed with the help of control groups.

Finally, the new list of technologies obtained, was compared to the list of technologies identified at the beginning with the analysis of twenty-five papers for removing duplicates, so the dictionary was ready to use.

On one hand, this analysis confirmed the relationship between Industry 4.0 and Precision Agriculture domains, on the other hand, it allowed to create a list of over 1000 technologies referring to the Precision Agriculture domain, by expanding the list generated thanks to the analysis of the 25 most cited papers on Precision Agriculture. This analysis shows how the intersection between the technologies belonging to Industry 4.0 and Precision Agriculture is very broad and makes the two concepts very close from a technological point of view. To investigate the presence of possible technological clusters, a text mining on abstracts of papers belonging to the "Precision Agriculture" domain was performed.

The block diagram in Figure 8.12 describes the process to create the dictionary.

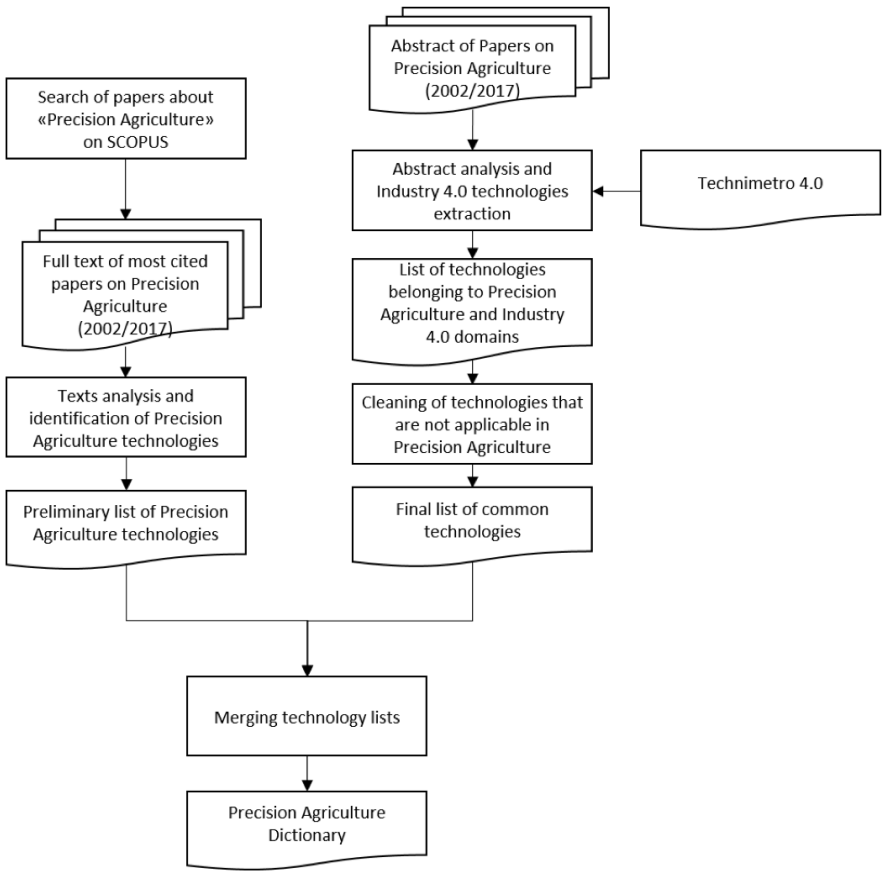

Figure 8.12: Process to create Precision Agriculture dictionary.

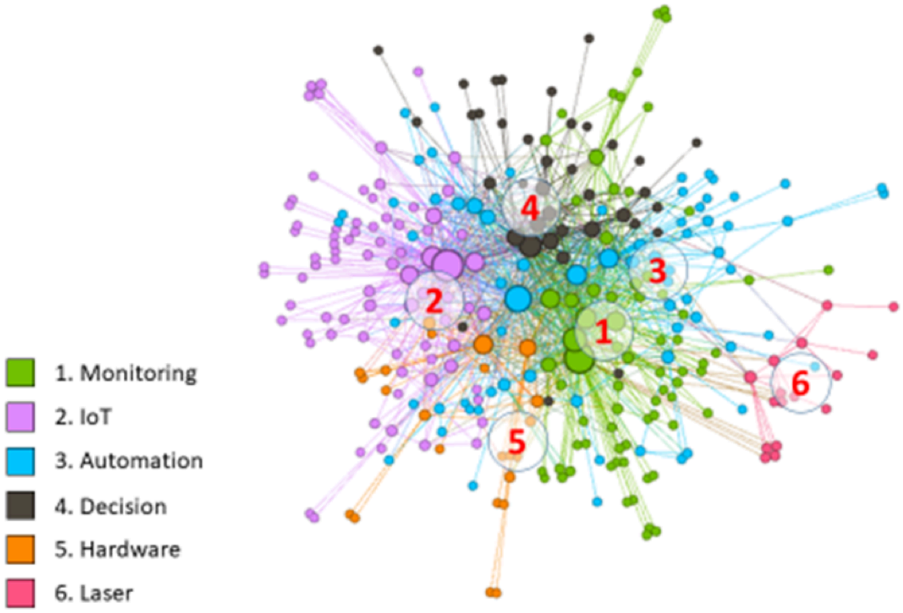

Figure 8.13: Graph of the Precion Agriculture technologies.

8.4.2 Results

The dictionary of Precision Agricultural Technologies includes 324 terms. Thanks to a graph, the most cited technologies and the connections between them allowed to identify at least 6 technology clusters. The graph in Figure 8.13 shows the structure of the dictionary containing the technologies related to Precision Agriculture and the relationships between the technologies that compose it. This representation allows to deepen the connections between the different clusters and technologies. The connections are represented by the lines that join the different nodes (which represent Precision Agriculture technologies) of the graph.

The size of the nodes varies proportionally to the number of papers they are cited in, instead their position depends on the number of connections between the different technologies: the most connected ones acquire a more central position into the graph and vice versa. A sample of the content of each cluster is:

- Monitoring: GPS, GIS, Data processing, GSM, Satellite, Ultrasound, Lidar, Broadband, Cellular, ...
- IoT: Wireless sensor network, Internet of things, RFID, Bluetooth, Zigbee, Wi-fi, Microcontroller, Arduino, ...
- Automation: Autonomous vehicle, Mobile Robot, Unmanned aerial vehicle, Agricultural robot, Computer vision, Data management, ...
- Decision: Artificial intelligence, Data mining, Expert systems, Forecasting, Machine learning, Semantic web, Smart grid, ...
- Hardware: Embedded system, Cyber-physical system, Manure spreader, Raspberry pi, CMOS, FPGA, ...
- Laser: Laser, Laser transmitter, Laser receiver, Laser surveying, Optical fiber, Photonic sensor, ...

Chapter 9

Wikipedia

Wikipedia is a non-conventional source of technical knowledge, with respect to patent and papers. With respect to the literature on field delineation and clustering applied to science and technology we innovate by introducing this new source. In particular in the present chapter, Wikipedia is proposed as tool to map a technological field. The process start with a small number of documents following a procedure which is *expert-independent*, in order to minimize the distortions from subjective judgment. We then exploit the properties of Wikipedia in order to delineate the field and identifying the linkages between technologies. Further description of the structure of Wikipedia can be found in section 6.3

In this section we present two methodologies capable of automatically use Wikipedia inforamtion and its structure with the aim of mapping and analysing this content. The results of the methodologies are described, togheter with example of applications of the extracted entities for intelligence tasks.

9.1 Industry 4.0: Extracting and Mapping Technologies

Industry 4.0 is getting the center of the scene with respect to the future of production systems in advanced countries and to its economic and social implications. It is considered as the new fundamental paradigm shift in industrial production. The new paradigm is based on the advanced digitalization of factories, the Internet, and future-oriented technologies bringing intelligence in devices, machines, and systems (Lasi et al., 2014). Despite its growing popularity and the great expectations in terms of innovation impact, the concept of Industry 4.0 remains strongly linked to technologies and frameworks that have been heavily researched and analyzed in the last decades. In particular, Industry 4.0 can be seen as a smart recombination of existing technologies and some new technologies and their application to the manufacturing environment (Trappey et al., 2016). This recombinant nature has led some authors to claim that it is nothing more than a re-labeling of old technologies, such as Computer Integrated Manufacturing (Apreda et al., 2016).

Yet other authors claim that this new wave of technology is fundamentally different from previous technologies and not just an amalgamation. In order to address the question whether Industry 4.0 is a new paradigm, or rather a re-labeling of existing technologies, a preliminary activity is needed, namely the delineation of the field and the clustering of technologies covered in the perimeter. It turns out that this activity is extremely challenging in the case of Industry 4.0, for a number of reasons we discuss in great detail. Faced with the complexity of Industry 4.0 existing delineation and clustering methodologies can be considered inadequate.

In this paper we develop a novel approach, test it, and show its superior performance with respect to other approaches. The key features of the approach are as follows:

i) the description of Industry 4.0 is offered in the form of an "enriched dictionary", or an ordered and comprehensive collection of lemmas, each of which are associated to full scale definitions and descriptions and to explicit linkages to other lemmas;

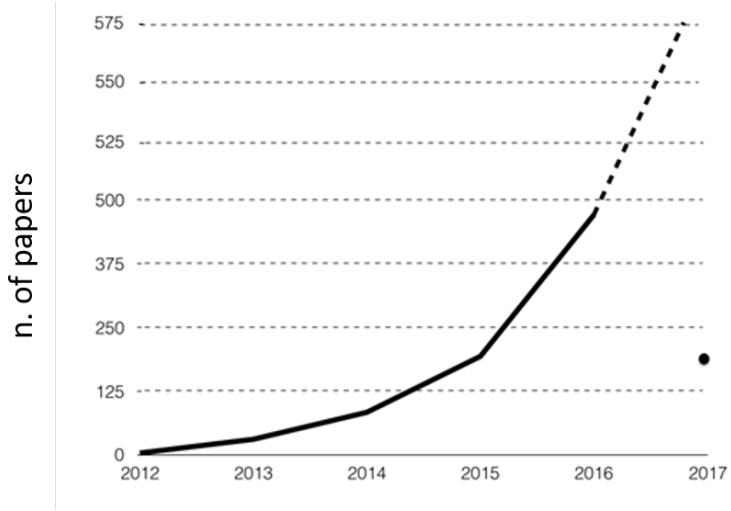

Figure 9.1: Trend of publications on Industry 4.0 (Title, Abstract, Keywords). Source: Scopus. Date: 06/06/2017

ii) the description of constituent technologies offered in the enriched dictionary is not obtained from individual experts, but is generated by accessing appropriate pages of the online encyclopedia Wikipedia;

iii) the total number of technologies covered is more than 1200, linked with more than 39,000 semantic relations;

iv) the perimeter of Industry 4.0 is not defined externally to the technology (by experts, government policies or other external sources) but is generated endogenously by examining the linkages between technologies described in the Wikipedia pages;

v) the update of the descriptions of technologies in the dictionary takes place in real time due to the distributed, parallel and selfcontrolled activities of authors in the worldwide community of contributors to Wikipedia;

vi) new technologies are automatically included in the dictionary if they exhibit a given threshold of connectivity with those already included in the perimeter.

Industry 4.0 as a multi-technology multi-stakeholder field

Industry 4.0 is the main keyword used by researchers, policy makers and entrepreneurs when describing how worldwide industrial systems will evolve in the near future by leveraging Internet connected technologies to generate new added value for organizations and society (Roblek et al., 2016). The growing interest is confirmed by the increasing number of academic papers focusing on topics that are related to the so-called "Fourth Industrial Revolution". As shown in Figure 9.1 the query "Industry 4.0" generates 967 papers. Even if the query is very sharp and does not include all the research efforts on the single "enabling technology" it demonstrates an exponential growth of the topic. In Figure 9.1 a projection represented by the dotted line is included. The projection has been drawn by considering a constant increase of the derivative calculated as the average of the last 4 years. Our forecast is that in 2017 there will be 575 new papers in Scopus; this estimate is supported by the fact that about 200 papers have been already published before June 2017 (represented by the point). Since previous analyses in Scopus demonstrate a delay between publication and loading of 5–6 months our forecast seems to confirm a growing interest on the topic.

Table 9.1 shows how the scientific production on Industry 4.0 is divided among the main research fields (multiple attributions are possible in Scopus). In particular, it is possible to identify field specific-technologies that refers just to one or few sectors/business areas, and general purpose technologies that can be exploited in several sectors/business areas.

Table: Breakdown of industry 4.0 papers per research field. Source: SCOPUS date: 06/06/2017

Subject Area	Number of Publications
Engineering	645
Computer Science	410
Business, Management and Accounting	185
Decision Science	134
Material Science	90
Mathematics	87
Chemistry	52
Physics And Astronomy	45
Social Sciences	34
Energy	30

Formulated initially in Germany in 2011, the Industry 4.0 paradigm has been quickly translated, adapted and reinterpreted in developed and developing countries. Table **??**tab:tabledocs40scopus) offers a compilation of official documents of governments, agencies and international organizations that in a few years after the initial formulation have embraced the concept. [FILIPPO, cerca articoli e cita, leva tabella]

Table: Technical document and Scientific articles used to create the Technology Seed List.

Author(s)	Name/title of the source
France Government	New Industrial France - Building France's industrial future
Geissbauer et al.	Industry 4.0: Building the digital enterprise - Global Industry Survey
German Trade & Invest	Industrie 4.0 - Smart manufacturing for the future
Heng, S.	Industry 4.0 - Upgrading of Germany's industrial capabilities on the horizon
National Intelligent Factories Cluster	Research andInnovation Roadmap
Rüßmann et al.	Industry 4.0 - The future of productivity and growth in manufacturing industries
Siemens	On the Way to Industrie 4.0 – The Digital Enterprise
Smit et al.	Industry 4.0 - Study for the ITRE Committee.
The Government Office for Science	The future of manufacturing: a new era of opportunity and challenge for the UK
Wee et al.	Industry 4.0 - How to navigate digitization of the manufacturing sector
Gorecky et al.	Human-machine-interaction in the industry 4.0 era

Despite this rapid and impressive convergence of interest (in itself a clear demonstration of the interdependence of policies across the world), there is no common ground in the definition and delineation of the field even if a first definition of the goal of industry 4.0 have been presented since 1998 (Council et al., 1998).

More precisely, while there is a reasonable convergence on the architectural definition of Industry 4.0, as defined in a relatively loose way, there is still considerable disagreement and misalignment with respect to constituent technologies (Riel et al., 2017; Smit et al., 2016; O'Halloran and Kvochko, 2015).

Furthermore, many constituent technologies are included in the definition of Industry 4.0, and hence described in these documents, from a variety of perspective that reflect mainly the huge variety of application domains. In other words, technologies are often described not only with respect to their fundamental engineering principles and related dimensions of performance, but with respect to specific applications to various manufacturing or service operations. In these applications the specific working of technologies and the associated dimensions of performance are indeed quite diverse.

Grangel and González (Grangel-González et al., 2016) develop a deductive rule-based system able to identify conflicts among AutomationML documents, named ALLIGATOR. It is interesting for the present work to notice how ALLIGATOR has the function to interoperate and align information models between a vast variety of areas (manufacturing, security, logistics) at a micro/plant level. In other words, this paper highlights the fact that one of the main problems of Industry 4.0 is the integration of models and concepts typically developed in their respective domains.

To offer an example of this state of affairs, let us consider the case of RFID (Radio Frequency Identification and Detection) technology. One of the main uses of RFID technology, that is, the detection of the location of a

Figure 9.2: A Porter-like Value chain framework for Industry 4.0 (courtesy of Towel Publishing).

tag moving along a known path with known speed can indeed be applied for largely different purposes (safety, tracking, localisation) and in various company areas (production, logistics, maintenance). In practice, each of these applications will develop the basic technology in different directions. Depending on the applications we will find largely different descriptions of the technology involved. Figure 9.2 offers a Value Chain-like representation of Industry 4.0, showing the wide range of applications of constituent technologies.

An interesting consequence of this state of affairs is that there is disagreement also at the higher level of government documents describing Industry 4.0 as the main object for innovation and industrial policies: when describing the main components of Industry 4.0 the French government uses 47 technologies, against 39 technologies for the Italian government.

Summing up, the recombinant nature of Industry 4.0 creates several interrelated problems for profiling and mapping: - the number of constituent technologies is very large - the description and performance of constituent technologies depend critically on the specific application, hence on the business function/company area affected - the stakeholders are located in several organizational positions - the technical progress is very fast, with many (even if not all) constituent technologies facing rapid changes in their nature and performance.

Faced with this situation, a traditional approach to profiling and mapping would require a massive effort of keyword definition. A number of experts would be recruited in order to offer a representation of the field from their disciplinary or industry perspective. Extensive domain knowledge would be mobilized in such a way to build up detailed yet comprehensive maps of technologies. Based on these maps, governments, statistical offices and international organizations would work for a few years in line in the effort to identify the trends of technologies, the most important actors, the shares of individual countries or regions in the global landscape. For sure, this is the approach underlying most of the documents illustrated in Table **??**tab:tabledocs40scopus) and this the approach that will be pursued in the years to come.

We strongly suggest this approach does not deliver the expected result. Keyword-based approaches to emerging technologies are too dependent on subjective judgments of experts. Even when the experts involved are top class and disinterested (often the best researchers or industrialists), their vision is inevitably partial. Even more importantly, keyword-based representations cannot be updated with the same speed of technology. The set of keywords identified by experts becomes inevitably obsolete in a few months. This paper leverages on publications and open source repositories to design, develop and test a methodology that provides delineation and clustering of technologies of the Industry 4.0 paradigm. The methodology is based on a dictionary concerning the enabling technologies for industry 4.0 with full definitions and connections between them. Given the fast growth and the uncertainty that characterizes industry 4.0 technologies, the present

methodology is designed to be a bottom-up and continuous evolving tool. The structure and measurements made on the tool refer to July 2017.

Mapping and Clustering a complex emerging technology

The first task for mapping a new technology is field delineation, or the definition of the perimeter of the field. Industry 4.0 is not a new technology, but a novel combination of partly existing, partly new technologies driven by the convergence of their trajectories. As a matter of fact, it is clearly an example of emerging technology (Rotolo et al., 2015), as it shares the features of rapid growth, technological uncertainty, and market uncertainty.

The delineation of new fields of science and technology is an issue addressed since the late '70s, after the pioneering period of bibliometrics. Field delineation is a necessary step when existing classifications do not offer timely, reliable or comprehensive coverage of a topics, for example of a new technology or a new technological field. Moving beyond existing classifications require undertaking a search which, in general, may follow a lexical approach, a citationist approach, or a mix between the two (Small, 2006; Kreuchauff and Korzinov, 2017). In all cases there is a need to initialize the process, i.e. to identify a set of elements that constitute the starting point for searching.

The main approach has been based on keywords, to be identified in various regions of documents (title, abstract, keywords, full text of an article; title, abstract, claims, full text of a patent) and to be used as queries. A query is a structured sequence of words, connected by logical elements, such as "or", "and" and the like, to be launched on a database in order to build up profiling, indexing, or clustering a given field. In general the initial set of keywords are provided by experts in the field, usually organized in an expert panel.

There are several limitations of the keyword/expert approach. First, expert based keyword definition, or patent classification is a very expensive activity (Tseng et al., 2007). Second, the keyword selection is based on subjective judgment, and when experts are asked to decide on relatedness measures (e.g. synonims, hypernims or hyponims), they do not apply systematic rules (Tseng et al., 2007, Noh et al. (2015)). Experts may be subject to a number of biases, such as for example the desirability bias (attributing higher probability of occurrence to preferred events) and many others (see section 5.8.2). Panels of experts are not immune by biases, such as group thinking. There is little research on the impact of expert subjective judgments on the delineation of emerging fields, but there is reason to believe it may be significant.

Third, the delineation of perimeter of emerging technologies is not robust to slight differences in the queries. As it has been shown by Zitt and Bassecoulard, little differences in the wording of queries, or on the time window, may end up in completely different sets of documents (Zitt and Bassecoulard, 2006). Therefore there is no proof that the method is reliable.

Finally, and more problematic for the case of Industry 4.0, the methodology is static, as it is based on a fixed sets of words. This set can (and in practice often is) updated, but this introduces a delay in the process and does not deliver reliable results. Keeping updated a collection of keywords in a dynamic technological landscape is extremely difficult.

These limitations have become evident in the last two decades, after the efforts of many authors to produce reliable perimeters of the emerging field of Nanotechnology. The initial efforts have been based on a classical expert-based approach. Panel of experts provided lists of keywords that were transformed into database queries. Among them, a consulting company called Lux, the Fraunhofer ISI in Germany, and CWTS in the Netherlands were the most active. Most studies delivered largely different delineations (Youtie et al., 2008; Ghazinoory et al., 2013; Ozcan and Islam, 2017).

In turn, these limitations opened the way to massive efforts to reduce the dependence on experts and exploit systematically the new opportunities opened by text mining, following what has been called "full text based scientometrics" (Boyack et al., 2013). Starting from the late '90 s several attempts have been made to apply text mining techniques to the patent corpus and the field is currently burgeoning (Joung and Kim, 2017; Ozcan and Islam, 2017).

Summing up, the existing approaches to the delineation of emerging technologies, taken together, suffer from the following limitations: 1. dependence on expert judgment 2. lack of robustness to alternative definitions of the perimeter 3. delay in update of technologies (static approach) iv) lack of transparency in the modeling of technology.

We now turn to the second main task in text mining of technology documents, that is, *clustering*. Once the perimeter of the field is delineated, a task usually included in the mapping is clustering, or the creation of groups of entities in such a way to reduce the complexity of the representation. Within text mining the clustering of documents is based on various kinds of linkages that are considered a signal of similarity in topics (see section 5.5.4). In general, linkages among documents can be generated by citations (citationbased clustering) or by the extraction of features in texts (text-based clustering) (Leydesdorff and Hellsten, 2006; Wang and Koopman, 2017; Jaffe et al., 1993).

The most used approaches to clustering assume that members of a cluster:

– cite each other (citation analysis: (Jaffe et al., 1993; Moed, 2006; Verspagen, 2007; Lee and Kim, 2017));

– share certain words (co-word analysis: (Callon et al., 1983; Rip and Courtial, 1984; Leydesdroff, 1989; Engelsman and van Raan, 1994; Van Raan and Tijssen, 1993; Yoon and Park, 2004));

– share a reference in their bibliography (bibliographic coupling: (Glänzel and Czerwon, 1996; Kuusi and Meyer, 2007));

– share the same sub-fields in a classification (co-subfield analysis:(Chang et al., 2009));

– are cited by the same documents (co-citation analysis: (Small, 1973; Small and Sweeney, 1985));

– are cited by the same authors (author co-citation analysis: (White and Griffith, 1981)).

In the proposed methodology we exploit the hyperlink feature of Wikipedia in order to introduce a new approach to clustering. Hyperlinks are introduced by authors in order to establish a semantic linkage between the page and other pages. We exploit this feature as follows: members of the same cluster are those pages that share the hyperlinks to other pages, according to thresholds defined by an appropriate algorithm. Note that hyperlinks are only superficially similar to citations. The origin page "cites" another page by hyperlinking it, but in effect this linkage is not a citation (that is, a reference to a previous work) but a signal of semantic similarity, intended to guide the reader in the network of meanings. We suggest that the hyperlinks are similar to citations under some respect, but very different under some other respects. Like citations, hyperlinks are introduced in the text by authors and reflect intentionality. Unlike citations, they reflect semantic relations, not relations of credit assignment or tribute to scientific authority. Perhaps more importantly, citations are introduced only by the author(s) of a paper and remain unchanged after publication. Hyperlinks, on the contrary, are introduced also by subsequent readers of the Wikipedia page. If the introduction of hyperlinks is not considered appropriate by the community of contributors, as it may happen due to vandalism, they are immediately removed (see the discussion below). This means that they reflect all possible semantic connections among the pages, as collectively stated by a large community of authors, in a reliable and robust way. We believe this methodology offers a remarkable improvement with respect to existing approaches.

How is similarity measured by using Wikipedia

In general, semantic relatedness is a measure of the similarity between two terms. It can be computed by statistical methods without requiring a manually encoded taxonomy, for example by analyzing term co-occurrence in a large corpus (Resnik, 1999; Jiang and Conrath, 1997). Wikipedia has been largely exploited in the literature in order to compute semantic relatedness. Gabrilovich and Markovitch (Gabrilovich and Markovitch, 2007) developed an alternative to Latent Semantic Analysis and called this new technique Explicit Semantic Analysis (ESA). This methodology first uses a classifier that is centroid-based to map input text to a vector of weighted Wikipedia articles. Then the vectors are exploited to obtain the semantic relatedness between two terms by computing the cosine similarity. This technique could be applicable to individual words, phrases or even entire documents. Furthermore, the mapping developed in this work has been successfully utilized for documents categorization. A new version of this kind of systems was

presented by Milne (Milne, 2007). While in Gabrilovich and Markovitch (Gabrilovich and Markovitch, 2007) the authors use the full text of Wikipedia articles to establish relatedness between terms, in this work only the internal hyperlinks are exploited. To compute the relatedness between two terms, they are first mapped to corresponding Wikipedia articles and then vectors are created containing the links to other Wikipedia articles that occur in these articles. The main problem facing semantic relatedness using Wikipedia is the disambiguation of terms. Several strategies have been developed to solve this problem. A first approach, described in Strube and Ponzetto (Strube and Ponzetto, 2006), exploits the order in which entries occur in the disambiguation pages of Wikipedia to find the most likely correct meaning. On the other hand, Gabrilovich and Markovitch (Gabrilovich and Markovitch, 2007) avoids disambiguation entirely by simultaneously associating a term with several Wikipedia articles. Milne (Milne, 2007) approach hinges upon correct mapping of terms to Wikipedia articles. However, when terms are manually disambiguated, it has been shown that the systems of semantic relatedness computation are more accurate than the systems of automatic disambiguation (Medelyan et al., 2009).

Summing up, a consistent literature in the field of computational linguistics and text mining supports the notion that the use of Wikipedia articles as a knowledge base is justified and promising.

9.1.1 Methodology

In this section we give evidence of the methodological steps undertaken to build up the enriched dictionary referred to Industry 4.0. The dictionary contains technologies related to the Industry 4.0 paradigm, each of which is associated to the full set of relations with other technologies. The dictionary is semi-automatically generated using technical documents and the Wikipedia free online encyclopedia. The linkages between technologies have all a semantic content, since they are generated within the text of the articles of Wikipedia when a related topic or entry is considered necessary for the logical flow of the definition or description.

Figure 9.3 shows the methodological steps needed to generate the enriched dictionary. In the flow diagram three elements are graphically displayed: activities (rectangular shape), check points (diamond shape) and documents created from the procedure (sheet of paper shape).

Generation of the seed list

As input for our methodology we used technical documents, official government documents and the most cited academic papers in the field of industry 4.0. The selection of seed documents has been made by:

– taking the official government documents of the US government and of three large European countries (France, Germany, UK), all of which are strongly committed to the support of Industry 4.0 and are widely considered a benchmark in international documents (for example, OECD and European Union), plus a selection of technical documents cited in government papers, for a total of 10 documents;

– taking the 10 most cited papers on Industry 4.0 according to Scopus. The reference to Industry 4.0 was explicit in the title, abstract and keywords of the papers. The extraction was made on Scopus Database in June 2017.

We deliberately limit ourselves to a small number of documents. The reason is explained in conjunction with the chosen stopping rules. The documents have been manually parsed by a team of Master students in Engineering Management at the University of Pisa. The assignment was “Understand each document and extract the enabling technologies for industry 4.0”. The manual search in the documents continued until the team reached the goal of 100 different technologies extracted. For our purposes 100 different technologies represents a reasonable seed list of technologies, which will be used as input for the automatic expansion phase. The chosen documents are shown in Table **??**tab:tabledocs40scopus).

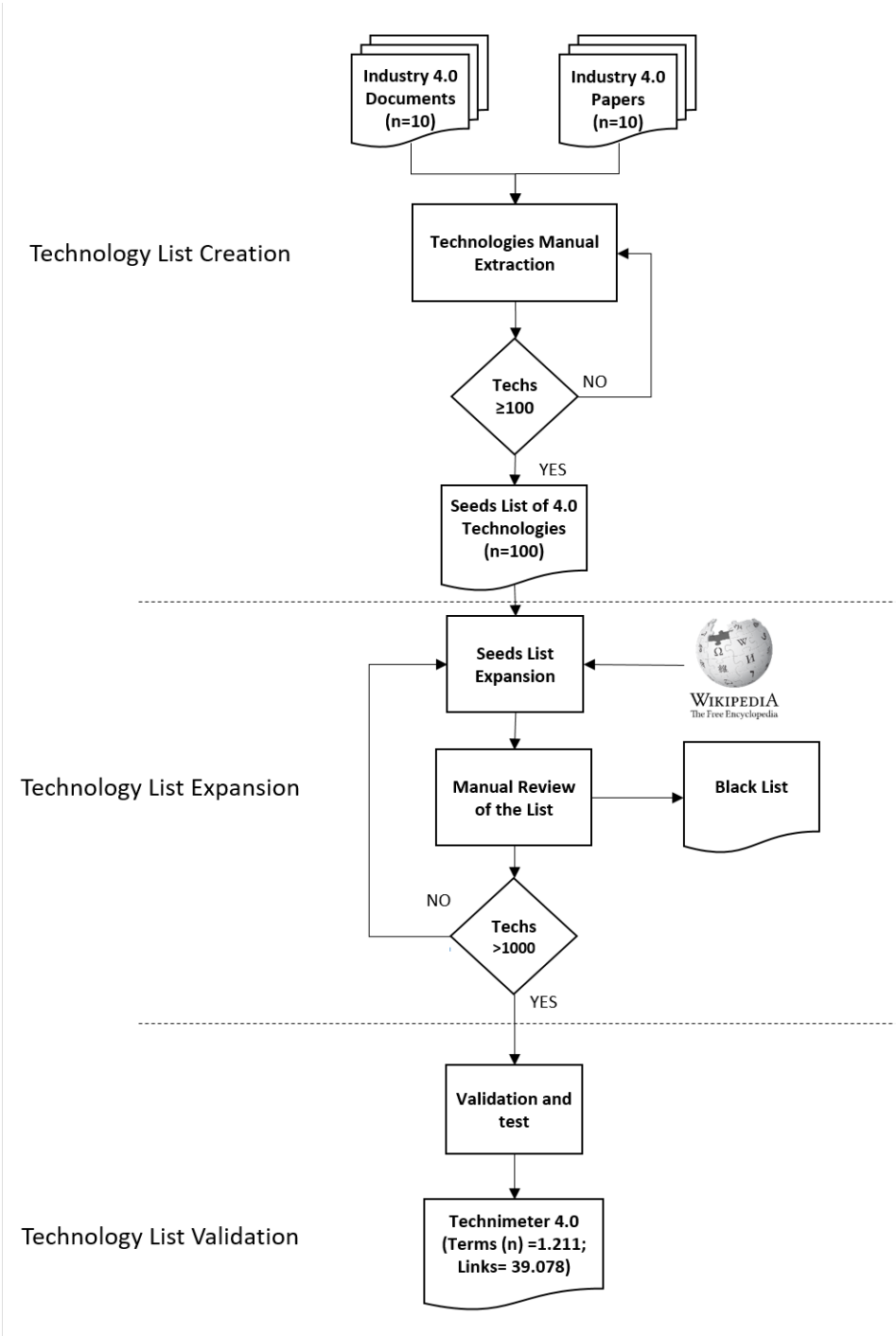

Figure 9.3: Flow diagram of the adopeted methodology.

Seed list expansion

For each term in the seed list the corresponding page in Wikipedia was found. All technologies identified in the seed were covered by Wikipedia. This is a preliminary confirmation of its relevance as a source of knowledge. These pages formed the initial glossary. The expansion procedure automatically retrieved the pages and identified all hyperlinks included in the description of the technology. The pages that are the target of hyperlinks are classified manually according to the following categorization:

a. links to pages already in the seed: these pages are labeled "anchors", since they provide robust indicators of technologies that are, at the same time, mentioned explicitly in the seed documents and referred to in Wikipedia pages that deal with other technologies;
b. links to pages not in the seed: these are labeled "missing technologies" and are stored in memory for later treatment as potential candidates to inclusion in the dictionary;
c. links to pages with non-technological content: they are labeled "stopwords" and are eliminated from the procedure. The overall procedure is iterated up to the point in which a number of at least 1000 different technologies is reached. At this point the procedure will stop. Updates and changes of the dictionary can originate from new entries (new technologies) or from updates to existing pages.

Given that the automatic extraction of Wikipedia pages can be run on a permanent basis, each version of the dictionary has a date. The current version, illustrated in the rest of the paper, is at the time point of July 15, 2017.

The stopping rule for the manual and the automatic expansion

The three stopping rules follow a sequential logic of order of mag- nitude: we start with approximately 101 documents, from which we extract 102 names of technologies, that, used as inputs to Wikipedia, deliver approximately 103 final technologies. More in detail we make use of:

1. 20 input documents: all technical documents taken as reference for Industry 4.0 share the same framework (i.e. DIN:SPEC 91345:2106). Many of the documents contains the same informations/terms/ technologies. Furthermore these documents are technology focused and therefore from a low number of documents we obtain a large number of technologies.

2. seed list of 100 technologies: as demonstrated in the past Wikipedia is a good source also for technical terms, therefore we assumed it was able to quickly expand the technologies related to Industry 4.0. For this reason the seed list is composed of no more than 102 entries.

3. output list of at least 1000 technologies: of the automatic expansion works correctly the technologies should increase by an order of magnitude in few iterations. Since the list has to be revised manually we decided for 103 entries as a target thus reducing the impact of manual review. As a matter of fact, the Wikipedia network of semantic linkages delivers a total number of technologies related to Industry 4.0 which exceeds this target, again confirming the validity of Wikipedia as a source of knowledge.

Structure of the enriched dictionary

The enriched dictionary can be defined as a set of enabling technologies for industry 4.0, associated to their definitions and to the linkages between them. The digital version of the tool is a hyperlinked text [1].

For the purpose of publication in an academic paper the tool can be represented as a table in which we have:

- Column 1- Technologies: Enabling technologies for industry 4.0, or a broad categorisation of technologies following a clustering procedure (see below)
- Column 2- Url: Links of the Wikipedia pages of the enabling tech- nologies for industry 4.0
- Column 3- Definitions: Glossary, snippets from wikipedia page of the definition of the enabling technologies for industry 4.0

[1] www.industria40senzaslogan.it

Enabling technology (A)	Wikipedia Page (B)	Definition (sample) (C)	Links (sample) (D)	Anchors (sample) (E)
3D printing	en.wikipedia.org/wiki/3D printing	3D printing, also known as additive manufacturing (AM), refers to processes used to create a three-dimensional object in which layers of material are formed under computer control to create an object. […]	3D bioprinting; Actuator; Artificial brain; Artificial muscle; Bikini; Biotechnology; Blue Brain Project; CAD; Delivery drone; Forbes; Gene therapy; Injection moulding; Inkjet; Laser-powered; Phosphor display;Magnetic refrigeration	Actuator; Artificial intelligence; Artificial photosynthesis; Atomtronics; Biotechnology; CNC; Computer-aided design; DMOZ; Electron beam melting; Home automation; Number; Internet of Things
Additive Manufactu Ring	en.wikipedia.org/wiki/**3D printing**	*(same as 3D printing)*	*(same as 3D printing)*	*(same as 3D printing)*
Augmented Reality	en.wikipedia.org/wiki/ Augmented_reality	Augmented reality (AR) is a live direct or indirect view of a physical, real-world environment whose elements are augmented (or supplemented) by computer-generated sensory input such as sound, video, graphics or GPS data. It is related to a more general concept called computer-mediated reality, in which a view of reality is modified (possibly even diminished rather than augmented) by a computer. […]	360° video; ARQuake; ARTag; ARToolKit; Ableton Live Accelerometer Acrobatics; Adobe Flash; Alexis Ohanian AlloSphere; Alternate reality game; American football; Android (operating system); Artificial reality; Audient; Augment (app) Augmented Reality Markup Language Augmented browsing; Augmented reality-based testing; Augmented virtuality Augmented web; Australia Council for the Arts; Automotive head-up display; Automotive navigation system; BBC; Bionic contact lens Blair MacIntyre; Blippar; Blob detection; Brain in a vat; Bruce H. Thomas; […]	Android (operating system); Bionic contact lens; Computer vision; Cyborg; DMOZ; EyeTap; GPS; Gesture recognition; Global Positioning System; Google Glass; Graphical user interface; Holography; IOS; MEMS; Mobile computing; QR code; RFID; Real-time computing Smartphone; Speech recognition; Tablet computer; Ubiquitous computing; Ultrasound; Virtual Reality; Virtual retinal display XML
Virtual Reality	en.wikipedia.org/wiki/ Virtual_reality	Virtual reality (VR) typically refers to computer technologies that use virtual reality headsets, sometimes in combination with physical spaces or multi-projected environments, to generate realistic images, sounds and other sensations that simulates a user's physical presence in a virtual or imaginary environment. […]	360° video; 3D audio effect; 3D computer graphics; A-ha; ACM Computing Classification System ARToolKit; Air force Algorithm; Algorithm design; AlloSphere; Alpenexpress; Alton Towers; Amnesty International Amusement arcade; Analysis of algorithms; Anshe Chung; Antonin Artaud; Anxiety disorder; Apple Inc.; Application security; Arcade game; [.]	3D computer graphics; ACM Computing Classification System Algorithm; Artificial intelligence Augmented reality; Computer; Computer-aided design Computer network; Computer security Computer vision; Computing platfor; Cryptography; Cyberspace; Data mining; […]

Figure 9.4: Sample of enabling technologies showing the table structure of dictionary.

- Column 4 Links: hyperlinks to other wikipedia pages from the wi- kipedia pages of enabling technologies for industry 4.0
- Column 5- Anchors: hyperlinks to other wikipedia pages that are enabling technologies for industry 4.0 from the wikipedia pages of enabling technologies for industry 4.

In the table shown in figure 9.4 shows a sample of the dictionary for four enabling technologies. An evidence is that there are conflicting definitions of "Augmented reality" (AR) and "Virtual reality" (VR); the first says that AR contrasts VR, while the second states that "AR systems may also be considered a form of VR". This is an evidence of the ambiguity that exists in the definition of 4.0 technologies. Moreover, the table underlines how words such as "3D printing" and "Additive manufacturing" that are used in different ways within papers and technical documents (see Table 3), basically refer to the same concept. Table 9.4 also shows the difference between links and anchors. Links include all hyperlinks found in Wikipedia pages. By nature, a certain fraction of these links contain information that is not relevant to our task. It is likely that, in the course of discussion of a given topics, authors quote an author (e.g. Bob Sproull in Virtual Reality), an institution (e.g. British Museum or California Institute of Technology) or an event or application (e.g. Coachella Valley Music and Arts Festival). These links do not add to our knowledge of the Industry 4.0 field. It can be seen that, following the definition of anchor given above, these terms are eliminated in column (E), which includes only the anchors, or those entries that are added to the body of knowledge.

9.1.2 Results

The dictionary is composed of 1.211 terms and 39.078 relationships between them. This generates a graph in which the node represents a technology and the edge represents a link in the Wikipedia page. The network structure naturally gives origin to graph-theoretic metrics. We exploit this property in order to generate a number of indicators that the readers may find it useful to examine.

Graph analysis and Sub-graph selection

Figure 9.5 gives an overview of the obtained. We compute for each node the in-degree (horizontal axis), the out-degree (vertical axis) and the number of links to other Wikipedia pages that each node has (color of the node). Our results show that in terms of the analyzed variables we can identify 4 different clusters of nodes (or technologies) each one having a different behaviour. The first group we take in consideration are the point having an indegree greater than 70 and an out-degree greater than 70. In this group on the top right of the page we have some outliers. These are terms like Microprocessor and Microcontroller, X86, 8-bit, 16-bit and 32-bit. Then we observe a sub-group centered on the coordinates (150, 100). Here we have terms

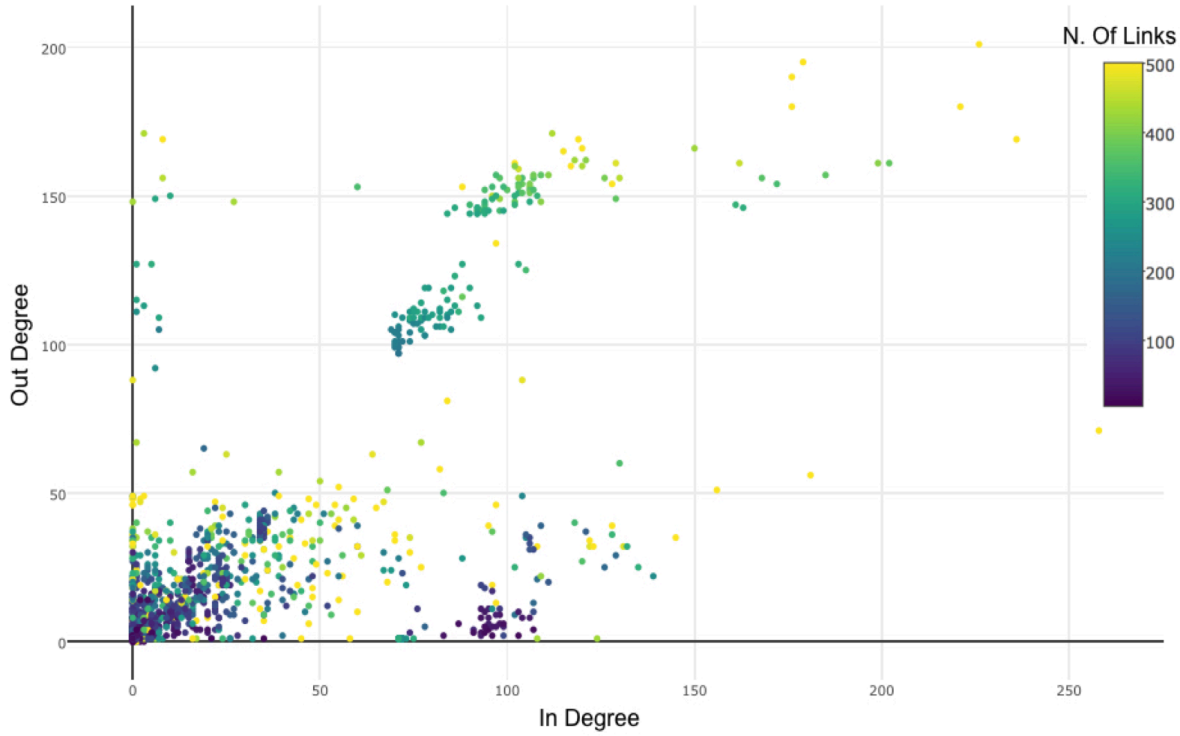

Figure 9.5: Plot of the in-degree and the out-degree of the nodes of the graph. The colour of the node represents the number of Wikipedia internal links.

like Program counter, Adressing mode, Instruction Cycle, Coprocessor, Symultaneus multitrading, SISD and MISD. Still within the first cluster we have another sub-group of terms centered in (80,110) with terms like Intel i860, Intel Atom, Intel 80286, Intel 80188 but also Bloomfiled and Wolfdale. The second cluster is one in which the points have an in-degree greater than 70 and an out-degree smaller than 70. Here we have terms like Machine learning, Artificial Neural Network, Cognitive Computing, Software, Random Access Memory, Internet, Firmware and C++. The third group collects points having an in-degree lower than 70 and an out-degree greater than 70. Here we have terms like Processors, Micro-Operation, Micro-Assembler, Application specific integrated circuit. The last and most populated cluster has an in-degree smaller than 70 and an out degree smaller than 70.

This cluster is more precisely visualized in Figure 9.6, for which we have as before the in-degree (horizontal axis), the out-degree (vertical axis) and the number of links to other Wikipedia pages that each node has (colour of the node). A jitter process has been implemented to the points on the graph in order to better visualize the overlapped points. In this plot only the technologies having both an in-degree and an outdegree lower than 70 are shown. This generates a subgraph composed by 931 nodes and 10673 edges.

Graph representation and cluster analysis

The structure of the graph offers it self naturally to clustering of technologies in order to obtain a readable mapping. The clustering algorithm receives as input the collection of technology terms T of the analyzed subgraph and returns a set of terms clusters C = {C1, C2,.., Cn} that cover the whole subgraph in analysis. Each cluster Ci is a subset of terms of T, and a term may belong to only one cluster. In Figure 9.7 we show a representation of the sub-graph made using Gephi software with its implementation of the Force Atlas algorithm (Bastian et al., 2009) . In this representation, two nodes in the graph are represented closely if they share an edge. In this way also nodes that belongs to the same communities of nodes (nodes that can be grouped into sets such that each set is densely connected internally) but do not share any edge are

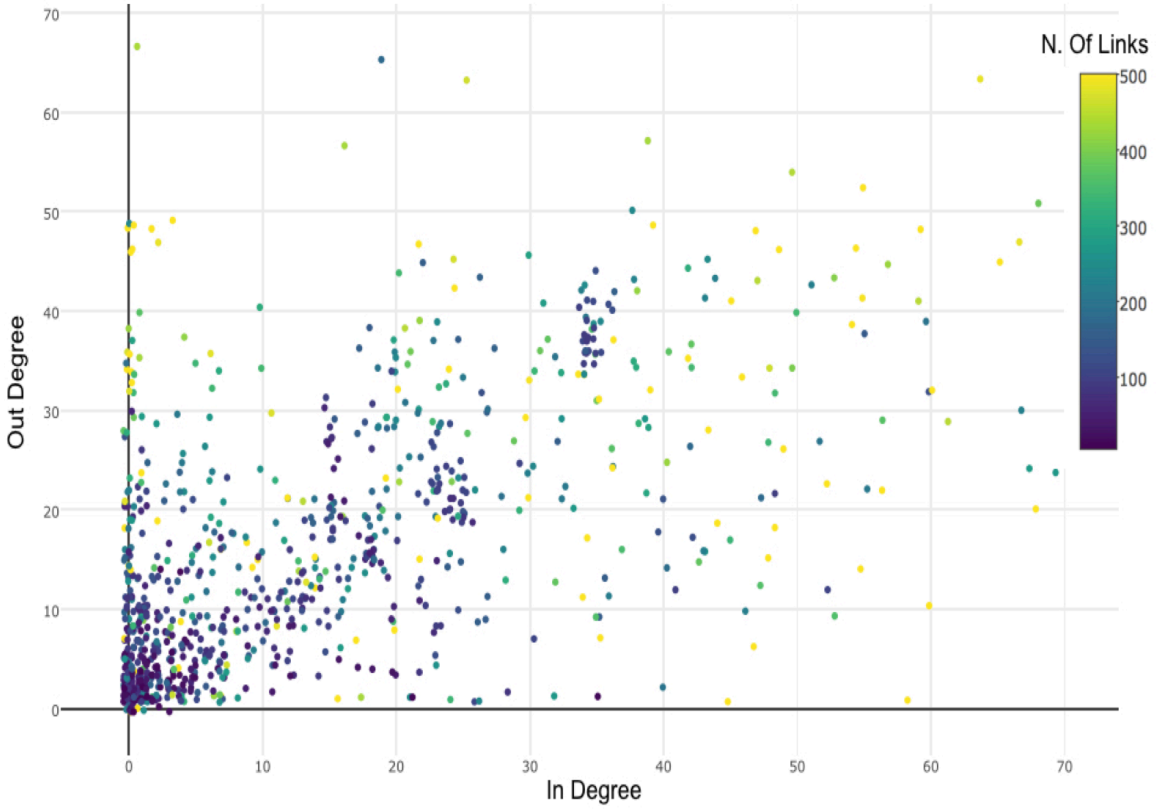

Figure 9.6: Plot of the in-degree and the out-degree of the nodes of the sub-graph. Selection of the nodes having an in-degree and an out-degree lower than 70. The colour of the node represent the number of Wikipedia internal links.

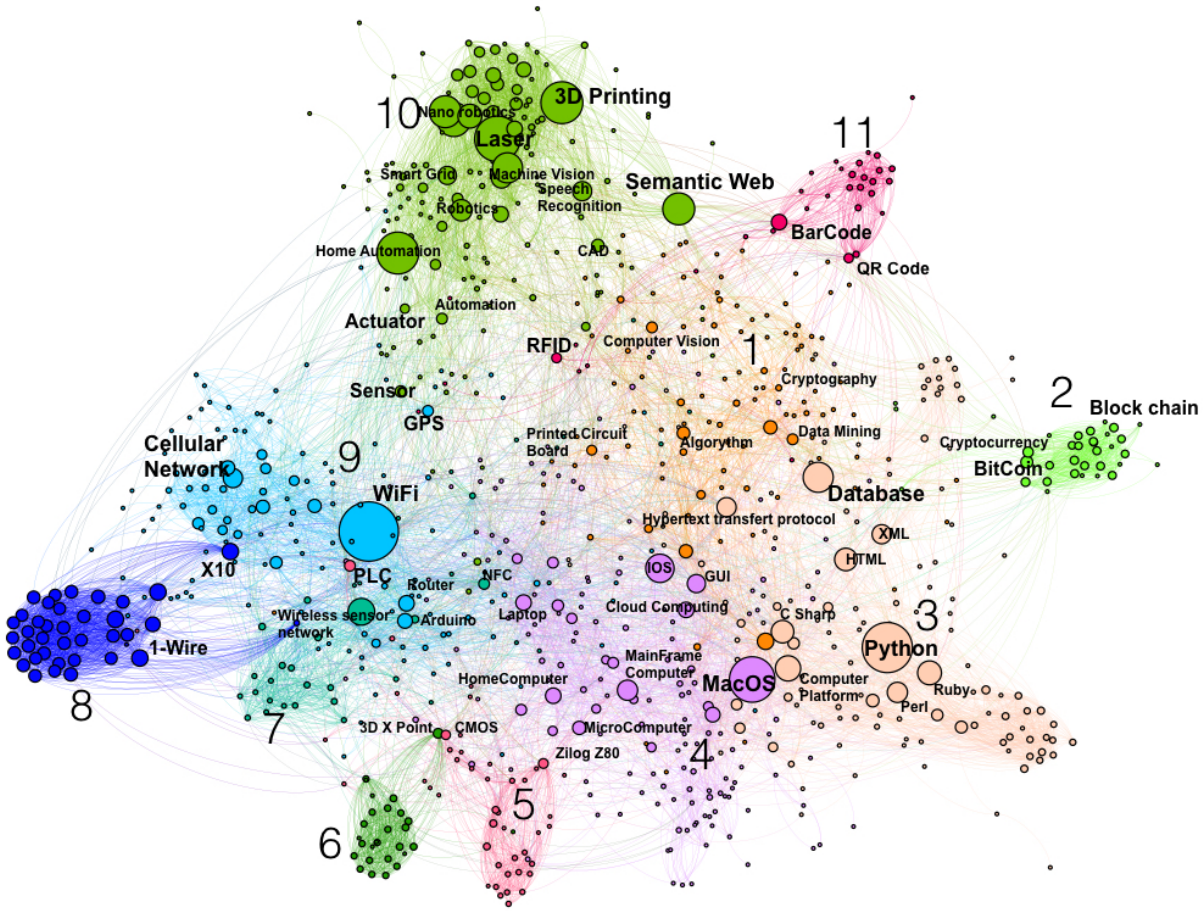

Figure 9.7: Representation of the graph of 4.0 technologies and of the clusters in wich them are arranged.

represented closely. In other words, the visualization tends to be coherent with the clustering algorithm. The size of the node is proportional to its in-degree while the colour express the cluster to which each node belong. Finally some of the labels of nodes and clusters are shown.

Each node is a technology and each edge represent a Wikipedia link between the pages. The size of the nodes is proportional to the in-degree, and the colours represent the clusters.

The algorithm we used to compute the modularity of each node and thus to assign a group to each of them is described in Blondel et al. (Blondel et al., 2008). The process resulted in 11 clusters. The content of each cluster is (for each cluster we can see the first 15 nodes in terms of in-degree):

1. *Big Data*: Virtual machine, Data mining, User interface, Algorithm, Computer vision, Cryptography, Printed circuit board, Middleware, Real-time computing, Virtual reality, Augmented reality, Human–computer interaction, Multiprocessing, Decision support system, Supervised learning
2. *Transactions, digital certification, digital currency*: Bitcoin, Cryptocurrency, Bitcoin network, Cryptocurrency tumbler, Digital currency exchanger, Alternative currency, Dogecoin, Ethereum, Litecoin, Monero (cryptocurrency), Namecoin, Peercoin, Virtual currency, Auroracoin, Blockchain, Lisk, Primecoin, Ripple (payment protocol), Titcoin, Zerocoin
3. *Programming languages*: Python (programming language), Database, Computing platform, Ruby (programming language), C Sharp (programming language), HTML, Perl, Hypertext Transfer Protocol, XML, Java (software platform), Haskell (programming language), .NET Framework, Lua (programming language), Sun Microsystems, BASIC

Figure 9.8: Wordcloud of the words belonging to the class 11 labelled as Identification. In this figure the size is proportional to the logarithm of the in- degree of each node that represents a word.

4. *Computing*: MacOS, IOS, Mainframe computer, Graphical user interface, Cloud computing, Home computer, Laptop, Solaris (operating system), Microcomputer, Personal digital assistant, QNX, Read-only memory, Tablet computer, ASCII, DOS
5. *Embedded Systems*: Programmable logic controller, Zilog Z80, CMOS, Zilog Z8, Toshiba TLCS, Zilog eZ80, NEC µPD780C, MOS Technology 6502, R800 (CPU), U880, Zilog Z180, Zilog Z800, Zilog Z8000, 1858 1, Hitachi HD64180
6. *Intel*: 3D XPoint, Intel ADX, Intel Clear Video, Intel SHA extensions, Intel System Development Kit, Intel 1103, Intel AZ210, Intel Cluster Ready, Intel Compute Stick, Intel Display Power Saving Technology, Intel Mobile Communications, Intel Modular Server System, Intel PRO/Wireless, Intel Quick Sync Video
7. *Internet of Things*: Wireless sensor network, Near field communication, Arduino, NetSim, Z-Wave, OPNET, Telemetry, RIOT (operating system), Routing protocol, TinyOS, Internet of things, NesC, MiWi, Nano-RK, LinuxMCE
8. *Protocols & Architectures*: 1-Wire, Profibus, Smart meter, X10 (industry standard), Modbus, Local Interconnect Network, TTEthernet, Fleet Management System, Keyword Protocol 2000, Meter-Bus, MTConnect, OPC Unified Architecture, PROFINET, RAPIEnet, SAE J1587
9. *Communication Network and Infrastructures*: Wi-Fi, Cellular network, Router (computing), Internet Protocol, Radiotelephone, ARPANET, Radio frequency, Digital subscriber line, General Packet Radio Service, Global Positioning System, CYCLADES, Beacon, Wireless, High Speed Packet Access, Evolution-Data Optimized
10. *Production*: Laser, 3D printing, Home automation, Agricultural robot, Nanorobotics, Semantic Web, Machine vision, Nanotechnology, Robotics, Information and communications technology, Speech recognition, Smart grid, Memristor, OLED, Computer-generated holography
11. *Identification*: Barcode, RFID, QR code, MaxiCode, Mobile tagging, Code 128, GS1 DataBar, High Capacity Color Barcode, Aztec Code, Barcode printer, Bokode, Codabar, CPC Binary Barcode, Interleaved 2 of 5, ITF-14

Let us examine more in depth a cluster, such as Identification (cluster 11). The size and colour of the words are proportional to the in-degree of the nodes as shown in Figure 9.8.

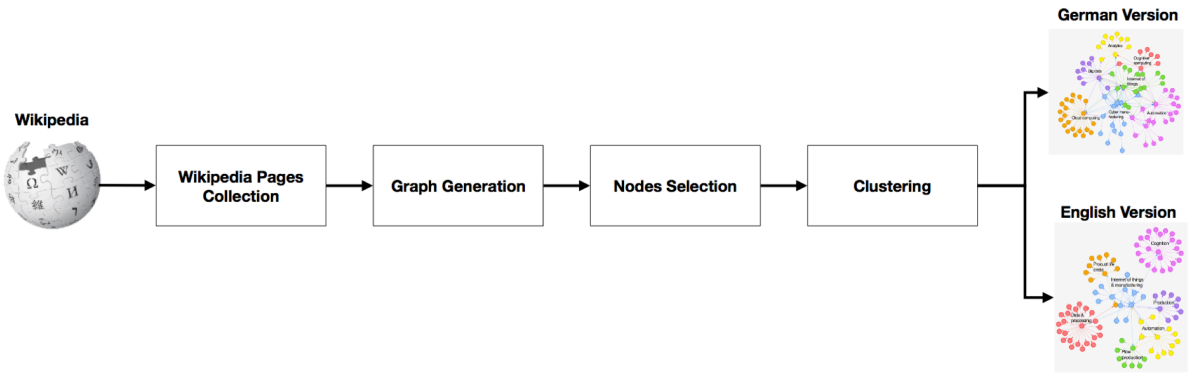

Figure 9.9: The workflow shows the steps we followed to generate the two graphs of industry 4.0 related concepts.

9.2 Industry 4.0: a Comparison with Industrie 4.0

The Industry 4.0 paradygim can be understood and implemented differently in different industrial ecosystems, depending on the country in which it is adopted. In the present section we make use of Wikipedia to build a network of pages regarding industry 4.0 in such a way taht it is possible to map how a country is actually implementing the industry 4.0 paradigm. To analyze the differences between the industry 4.0 concept in the country of origin (Germany) and the rest of the world, our approach is thus to investigate the Wikipedia pages of Industrie 4.0 and Industry 4.0 and the pages they are connected to. Using clustering algorithm it is then possible to explore the different focus that countries have on different technological sub-field of Industry 4.0.

All aspects during our analyses and the description of the project, e.g. page names, consist of two parts: one in English and one in German. For simplicity and better understandability, we well mostly refer to everything only by the English name.

9.2.1 Methodology

In the present section we explain the methodology we adopted to build and analyze the two graphs of Wikipedia pages representing the concepts linked to industry 4.0 in Germany and in the rest of the world. As a proxy of the world-wide vision of industry 4.0 we analyzed the English version of Wikipedia. The main steps of our methods are represented in the workflow of figure 9.9 The first step consisted of extracting all pages of interest from Wikipedia. Then the pages and their connections were used to build two graphs (the German and English version). The nodes of the graphs were then cleaned using a novel approach based on the categories of the Wikipedia pages. Finally a clustering algorithm was used to illustrate communities within the graph. These steps are described in the following sections.

Wikipedia Pages Collection

To list the pages connected to the concept of Industry 4.0, all Wikipedia pages that are linked to the page of Industry 4.0 (and Industrie 4.0) were downloaded. We will refer to these pages as level 1 connection. Then all the pages linked to the level 1 pages are collected and marked as level 2 pages. This process can be continued until a wanted level is reached: for the purposes of the present work we decided to show the effectiveness of the process stopping at level two. For each page, we collected the links to other pages and the categoriess of the page.

Graph Generation

Two pages that are linked by a link within Wikipedia shares a certain relations of which we ignore the meaning. Thus, it is possible that the approach we described until now, also collects pages that are not relevant for the analysis. In our case, this means that some of the extracted pages are not technologically related to the subject of Industry 4.0. Thus, the list of entries had to be cleaned from these non-relevant ones. To make the filtering process more efficient, we filtered considering the categories that had been extracted together with the pages. We considered as relevant all the categories of the pages of level 1. Then the level 2 pages were automatically considered to be relevant if they were in one of the formerly marked categories. For the nodes selection we also used a black-list of non-relevant Wikipedia pages (e.g. International Standard Name Identifier, International Standard Book Number, Digital object identifier, ISBN), previously developed by the authors in (see chapter 9.1). This list has been traduced in German for the purposes of the present work.

Clustering

In order to identify macro topics in the formerly created graphs, we applied clustering methods. The most beneficial results were found to be generated by applying a spin glass algorithm, which employs the spin glass model and uses simulated annealing to find communities in networks (Reichardt and Bornholdt, 2016). In this approach the community structure of the network is interpreted as the spin configuration that minimizes the energy of the spin glass with the spin states being the community indices.

9.2.2 Results

In this section we will focus on describing the differences that emerges between the two national technological approaches of industry 4.0 in Germany and in the rest of the world.

The collected pages

We collect 97 pages for the German graph and 95 pages for the English graph. Table 9.2.2 list a sample of forty pages per language, alphabetically ordered.

Table: For each IPC class the ratio of the percentage of patents containing at least one ® or ™ character is computed for two patents sets 1995 to 2002 and from 2007 to 2014.

German Version	English Version
Absatzwirtschaft	Adaptive system
Automatisierung	Agent-assisted automation
Automatisierungsgrad	Ambient intelligence
Automatisierungstechnik	AmbieSense
Autonomic Computing	Analytics
Autopoiesis	Artificial intelligence
Betriebswirtschaftslehre	Automated reasoning
Corporate Evolution	Automation
Cyber-physisches System	Autonomic computing
Dataset	Autonomous car
Daten	Big data
Datenabgleich	Big Data Maturity Model
Datenarchitektur	Big memory
Datenaustausch	Carrier cloud
Datenbasis	Cloud-based design and manufacturing
Datenelement	Cloud analytics
Datenfeld	Cloud collaboration
Datenmanagement	Cloud computing
Datenmodell	Cloud computing comparison
Datenmodellierung	Cloud computing security
Datensicherung	Cloud database
Datenstruktur	Cloud engineering
Datentyp	Cloud Foundry
Denken	Cloud management
Dienst (Informatik)	Cloud manufacturing
Dienstleistung	Cloud research
Digitales Objektgedächtnis	Cognitive computer
Elektronische Datenverarbeitung	Cognitive computing
Erfahrungskurve	Community cloud
Erinnerungsvermögen	Competitions and prizes in artificial intelligence
Feld (Datentyp)	Computer-aided engineering
Feldgerät	Computer-aided technologies
Fertigungslinie	Computer-integrated manufacturing
Fließbandabstimmung	Computer vision
Fließbandfertigung	Continuous analytics
Gegenwart	Control loop
Geschichte der Produktionstechnik	Control system
Glauben	Cyber-physical system
Gruppenfertigung	Cyber manufacturing
Halbzeug	Datafication

The Graphs

The dimension of the graphs in terms of nodes, links and categories are shown in Figure 9.10. From this figure is evident how the two graphs has almost the same number of nodes; by the way, the mean number of categories for the German version is close to 0.5 while its 0.3 for the English version. Beside these measure, as it would be even more evident from further analysis, there is the broader approach to industry 4.0 in the rest world with respect to the more focused vision that Germany has.

The resulting graphs are shown in figure 9.11 (respectively the German version on the left and the English version on the right). Here the vertices of the graphs represent the Wikipedia entries, while the links are depicted by arcs that are directed from the originating page to the linked page. The structures of the two

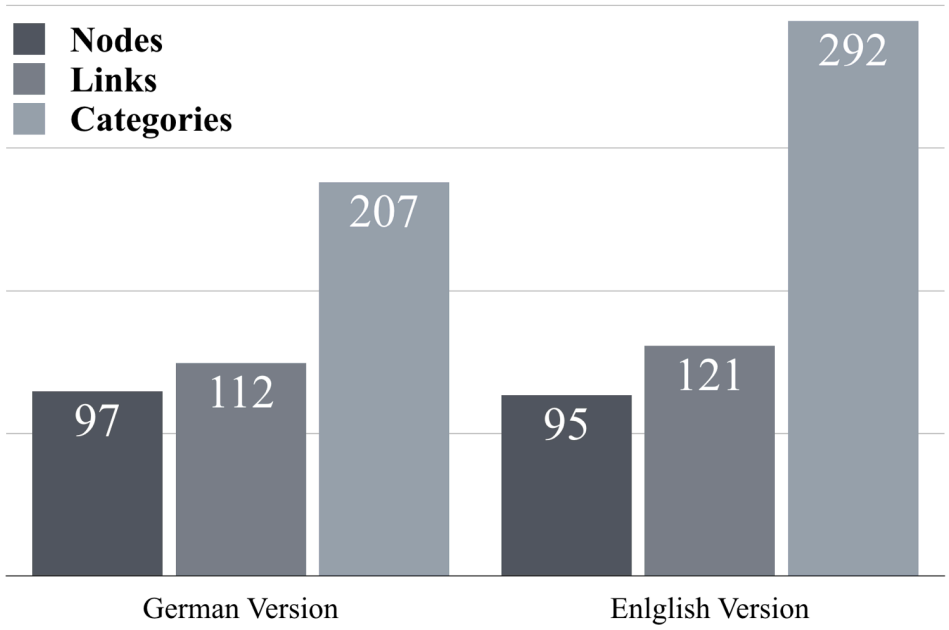

Figure 9.10: Number of nodes, links and categories for the German version of the graph (left) and the English version (right)

graphs and the difference in the number of links gives an evidence of the more focused network of concepts in Germany: the number of links is in fact lower for Germany, creating a more clustered graph.

Graph metrics

In this section we analyze the graphs using standard metrics. These measures allow investigations about relevant topology and specific features of a given graph. More precisely, we looked at the following measures: - Graph diameter: calculates the longest shortest path in a graph, thus, the minimal distance between two nodes that are the furthest away from each other. - Nodes Degree: Shows the number of nodes that every node is connected to. The mean value of the degree for the whole graph shows how interconnected the graph is: the higher the value, the more connections the nodes have with each other. - Nodes Distance: Gives back the minimum number of edges that connect one node with each other. Here again, the mean value shows the interconnectivity of the graph: A lower value means that the way from one node to another is shorter, thus that the graph is more highly connected.

The results of these metrics for the two graphs are shown in table 9.2.2. From the diameter measure we can see that all of the pages are connected through a maximum of 4 edges. This is not surprising, since it was a premise of the method when setting the level to 2. Considering the mean degree, a level of 2.38 for the German version and 3.29 for the English version shows how the nodes of the English graph are more interconnected. This can also be qualitative observed when looking at the figures 3. Finally, the average distance from one node to another is higher in the German graph, supporting the finding of the measures discussed above. These findings goes in line with what is said by a study of Acatech (Kagermann, 2006): in the USA, for example, the term Industry 4.0 is understood in a much wider meaning. Furthermore, Acatech study say that in Germany, the focus is mostly on technological dimensions, while in the USA, the development of the new business models in the area of big data analytics play a greater role, among others.

Table: For each IPC class the ratio of the percentage of patents containing at least one ® or ™ character is computed for two patents sets 1995 to 2002 and from 2007 to 2014.

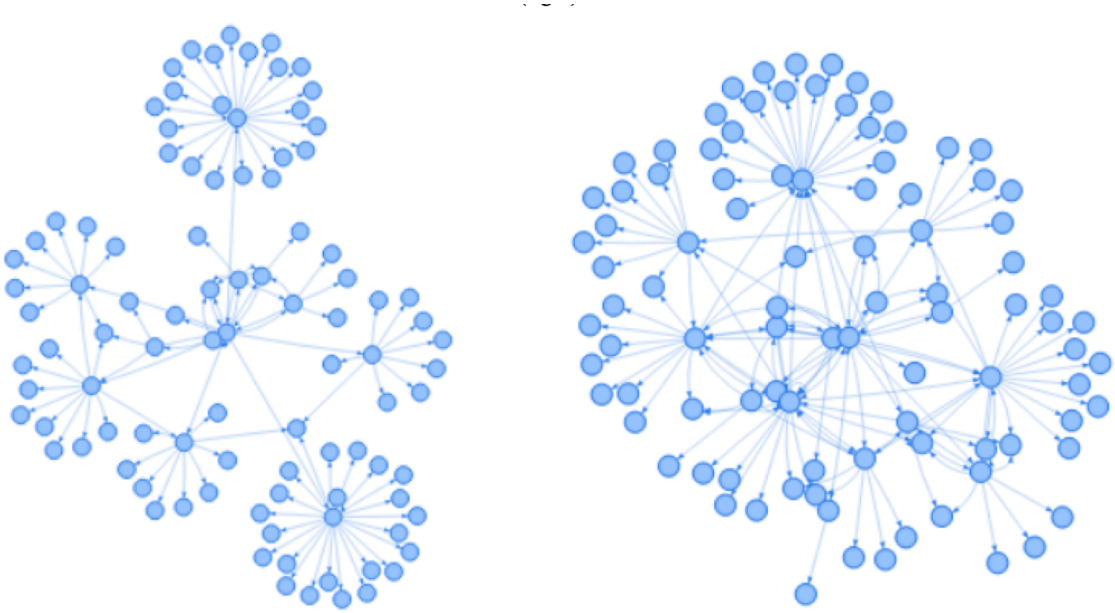

Figure 9.11: The German version (left) and the English version (right) of the Industry 4.0 related graphs.

Measures	German	English
Diameter	4	4
Mean Degree	2.38	3.29
Mean Distance	3.25	3

Clustering

Figure 9.12 shows the outcomes of the clustering process: the vertices of the graphs represent the Wikipedia pages, the links are depicted by arcs that are directed from the originating page to the linked page and the clusters are represented by the different colors of the nodes. In order to make these clusters comparable, the label of the German version are translated in English. The resulting, corresponding groups as well as the number of nodes they contain can be found in table 9.1.

Considering together figure 4 and table 3, it is evident how the central node of the graph(Industry 4.0) is contained in the first cluster Internet of things & manufacturing for both the graphs. It is also by design connected to all other clusters, so that no cluster is isolated from Industry 4.0 node. As evident from table 3, the clusters in the two graphs are similar. We have the same five clusters in both graphs and a different one for each language. This shows that Industry 4.0 framework if considered at a more abstract level, is similar in different countries. This is not surprising, since there must be something that made policy and industry call it the same name, but it is a strong result for our methods, able to map bottom-up this phenomenon.

However, when looking at the clusters more closely, some interesting differences get clear. The cluster production, product life circle & flow production, which is completely missing in the English graph, is the one with the higher number of nodes in the German one. This shows how the topic of Industry 4.0 in Germany is much more concentrated on production than elsewhere. In the study of acatech (Kagermann, 2006), it is also explicitly stated that in Germany " [...] the focus is on optimizing production processes in terms of quality, price and flexibility and delivering better financial returns overall". The focus on production for Germany is also confirmed by a monitor paper written by the by the European Commission (Comission, 2016). To give a clearer view of this phenomena, we have to consider that the word production is very similar in meaning to manufacturing. The cluster including the latter (internet of things & manufacturing) has more than the double number of nodes in the English graph with respect to the German one, so it might

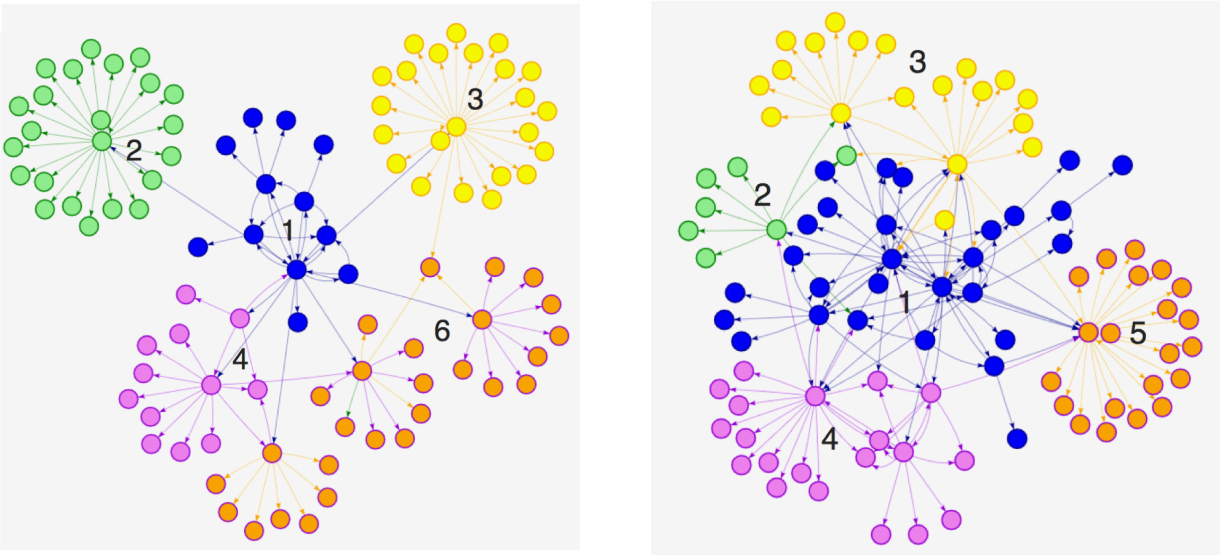

Figure 9.12: The German version (left) and the English version (right) of the Industry 4.0 clustered graphs.

be that it is rather the wording that makes the difference. Anyway, by looking at the the disposition of the cluster (see figure 9.12) internet of things & manufacturing in the two graphs, it can be observed that in the English version the cluster is strongly integrated with ITC-related topics (Data processing & Analytics, Cognition and Cloud Computing), adding the adjective cyber to the topic of manufacturing.

Table 9.1: Clusters names and number of nodes per cluster for each version of the Industry Graph 4.0.

Cluster	# Nodes German	# Nodes English
1- Internet of things & manufacturing	12	28
2- Cognition	23	7
3- Data, processing & analytics	23	19
4- Automation	12	20
5- Cloud computing	-	21
6- Production, product life circle and flow production	27	-

Additionally, also automation has more entries in the English version and is highly connected not only to internet of things & manufacturing, but also to could computing, which does not even exist in the German version of the graph: it is well known that Germany is not at the front when it comes to the integration of ICT in Industry 4.0. This found is again confirmed by the study of Acatech (Kagermann, 2006), where it is said that "Germany is currently lagging behind with regard to data-driven business models and the development of large platform ecosystems".

The fact that cloud computing does not show up in the German graph is examined more closely. In fact, a Wikipedia page about this topic does exist in the German version of Wikipedia, but is not linked to the Industrie 4.0 page. As stated in section 4.3, contributors of the German version of Wikipedia tend to add only the most important connections between pages. This again shows that integrated ICT is not the most important topic in Germany now.

Another great difference in the number of nodes is in the cluster cognition, which seems to be more important in the German version of the graph. The German Wikipedia page describes this as a processing and rearrangement of information between a human and a system. Acatech study (Kagermann, 2006) does the same observation: "In Germany in particular, the focus is on integrating information, communication and manufacturing technologies". This could explain even better the finding that the German graph lacks on the ICT side, giving a new interpretation: in Germany there is a different interpretation of the meaning of cognitive computing with respect to the rest of the world. Finally, while the group of data, processing and analytics nearly has the same amount of nodes, the links of this cluster with the other clusters is quite different in the two graphs and also its content changes from on nation to the other. Words like big data does not even exist in the list of German pages related to industry 4.0 where it is limited to just data.

Chapter 10

Social Media

The technical knowledge generated by Social Media remains largely untapped. Some researchers proved that social media have been a valuable source to predict the future outcomes of some events such as box-office movie revenues or political elections. Social media are nowadays also used by companies to measure the sentiment of customers about their brand and products.

This chpater proposes a new social media based model to measure how users perceive new products from a technical point of view. This model relies on the analysis of advantages and drawbacks of products, which are both important aspects evaluated by consumers during the buying decision process. This model is based on a lexicon developed in a related work (see chapter 7.2) to analyse patents and detect advantages and drawbacks connected to a certain technology.

The results show that when a product has a certain technological complexity and fuels a more technical debate, advantages and drawbacks analysis is more efficient than sentiment analysis in producing technical-functional judgements.

10.1 Technical Sentiment Analysis

Nowadays, social media have become an inseparable part of modern life, providing a vast record of mankind's everyday thoughts, feelings and actions. For this reason, there has been an increasing interest in research of exploiting social media as information source of knowledge although extracting a valuable signal is not a trivial task since social media data is noisy and must be filtered before proceeding with the analysis. In this domain, sentiment analysis, which aims to determine the sentiment content of a text unit, is considered one of the best data mining method. It relies on different approaches (Collomb et al., 2014) and it has been used to answer research questions in a variety of fields comprised the measure of customers perception of new products (Mirtalaie et al., 2018b). In this section, we try to understand if sentiment analysis is really the best available method to analyse consumer's perception of products, expecialy when we want to measure the perception of the technical content of the product. Thus we compare State of the art sentiment analysis techniques with a lexicon of advantages and drawbacks related to products.

10.1.1 Methodology

Our work started with the selection of an event able to polarise Twitter users' attention and products to analyse. In particular, we chose a premiere tradeshow for the video game industry, and two video game consoles disclosed during the event. We collected about 7 milions tweets about products published before, during and after the tradeshow. Since social media data is noisy (for example it may contains spam and advertising), before proceeding with the analyses, we filtered our dataset. In particular, after removing too short and non-English tweets, we manually classified a randomly extracted subset of posts to train a classifier

which provide us the cleansed dataset. Then we conducted a sentiment analysis of the tweets using state of the art machine learning techniques. We classified each tweet as positive, negative or neutral. At this point we applied our lexicon identifying advantages tweets and drawbacks tweets. Finally we compared the outputs of the two analyses for the two product-related clusters of tweets. We found consistent differences between the extractions. The results shows that when a product has a certain technological complexity and fuels a more technical debate, advantages and drawbacks analysis is more able than sentiment in producing technical-functional judgements. For this reason we think that the proposed methodology peforms better then standard sentiment analysis techniques when a product has a certain technological complexity and fuels a more technical social media discourse.

Selection of a triggering event and products

We chose the Electronic Entertainment Expo as event able to polarise users' attention. Commonly referred to as E3, it is a premier trade event for the video game industry, presented by the Entertainment Software Association (ESA). We chose two new video game consoles, disclosed at E3 2017, as products of which predicting the success or failure. The first is Xbox One X, a new high-end version of Xbox One with upgraded hardware and the other product is New Nintendo 2DS XL, a streamlined version of the handheld console New Nintendo 3DS XL.

Data collection

Twitter provides two possible ways to gather tweets: the Streaming Application Programming Interface (API) and the Search API. The first one allows user to obtain real-time access to tweets from an input query. The user first requests a connection to a stream of tweets from the server. Then, the server opens a streaming connections and tweets are streamed in as they occur, to the user. However, there are a few limitations of the Streaming API. First, language is not specifiable, resulting in a stream that contains tweets of all languages, including a few non-Latin-based alphabets, that complicates further analysis. Instead, Twitter Search API is a Representational State Transfer API which allows users to request specific queries of recent tweets. It allows filtering based on language, region, geolocation, and time. Unfortunately, using the Search API is expensive and there is a rate limit associated with the query. Because of these issues, we decided to go with the Twitter Streaming API instead. For each product, we detected related hashtags and keywords an constructed a query to download relevant tweets. We chose to collect tweets not only after the tradeshow, but also before. For these reason, we initially identified some products keywords with their provisional names and we updated them at a later stage. Tweets have been downloaded from CNR (Consiglio Nazionale delle Ricerche, Istituto di Informatica e Telematica, Area di Pisa) since 11th June 2017 h. 10:00 to 31th July 2017 h. 15:00.

Data filtering

The initial dataset resulted to be very noisy, containing tweets written in different languages, advertising and posts related to different products or subjects. We chose to keep into account only English tweets because sentiment and advantages/drawbacks lexicon is in this language. The data set is filtered removing tweets with less than five words and non-English posts with a language classifier. We obtained 7.165.216 of filtered tweets.
At this point we created a golden set of relevant tweet to train a Supported Vector Machine classifier able to recognize relevant and unrelevant tweets. We defined characteristics that make a tweet: (i) relevant (posted by users or containing words or opinions related to our products of interests and their functionalities), (ii) irrelevant (tweets containing advertisings, links to e-commerce websites or messages related to other products or subjects). A researcher manually classified a subset made up of randomly extracted tweets. In particular, we exctract a subset composed of 6.500 finding 105 positive results and 6.395 negative. SVM model was then trained using this dataset, and computed a probability for each tweet to be relevant or irrelevant. A threshold of 0.7 has been chosen to label a tweet as relevant. The final dataset of filtered tweets, made up

of 66.796 posts. We clustered tweets using product-related keywords. Clustering posts allowed us to further filter the final dataset which contained a small number of irrelevant tweets (Table 10.1).

Table 10.1: Clusters of tweets.

	N° of tweets	% of tweets
Xbox One X	64885	0.9714
New N2DS	1706	0.0255
Irrelevant tweets	198	0.0030

Sentiment analysis

Table 10.2 presents the results of the sentiment analysis. We classified each tweet according to its sentiment into positive, negative, or neutral. We used an established methodology developed by (Cimino et al., 2014). We pre-processed the tweets by removing mentions (@ character), URLs, product hashtags, emoticons and single characters. As a result, for each tweet we obtained a probability of belonging to a mood class. After a manual analysis, we used a class prediction probability threshold of 0.6 to filter out low confidence prediction, i.e. tweets that cannot be classified as positive or negative with a high confidence are classified as neutral instead.

Table 10.2: Clusters of tweets.

	Positive	Negative	Neutral
Xbox One X	0.3599	0.0465	0.5937
New N2DS	0.5299	0.0158	0.4543
Overall	0.3642	0.0457	0.5901

Advantages and drawbacks analysis

To extract technical advantages and drawbacks from tweets we used the lexicon developed in chapter 7.2 that contains 657 Advantages words and 297 Drawbacks clues. These words are searched on our dataset finding different percentages of tweets with words from the lexicon in the two product-related clusters of tweets. Table 10.3 reports the results.

Table 10.3: Percentages of tweets containing or not words from our lexicon.

	Adv	Drw	Adv & Drw	No Adv or Drw	Adv or Drw
Xbox One X	0.0884	0.0374	0.0037	0.8705	0.1295
New N2DS XL	0.0662	0.0094	0.0000	0.9244	0.0756

10.1.2 Results

We adapted the advantages & drawbacks analysis to give as output a classification ef each tweet. We classified data coming from the latter analysis in this way:

1. positive (tweets containing only advantages words)
2. negative (tweets containing only drawbacks)
3. neutral (tweets with no words of our lexicon or controversial tweets)

As we can see in figure 10.1, sentiment analysis is more able to polarise tweets. In fact, with this analysis we

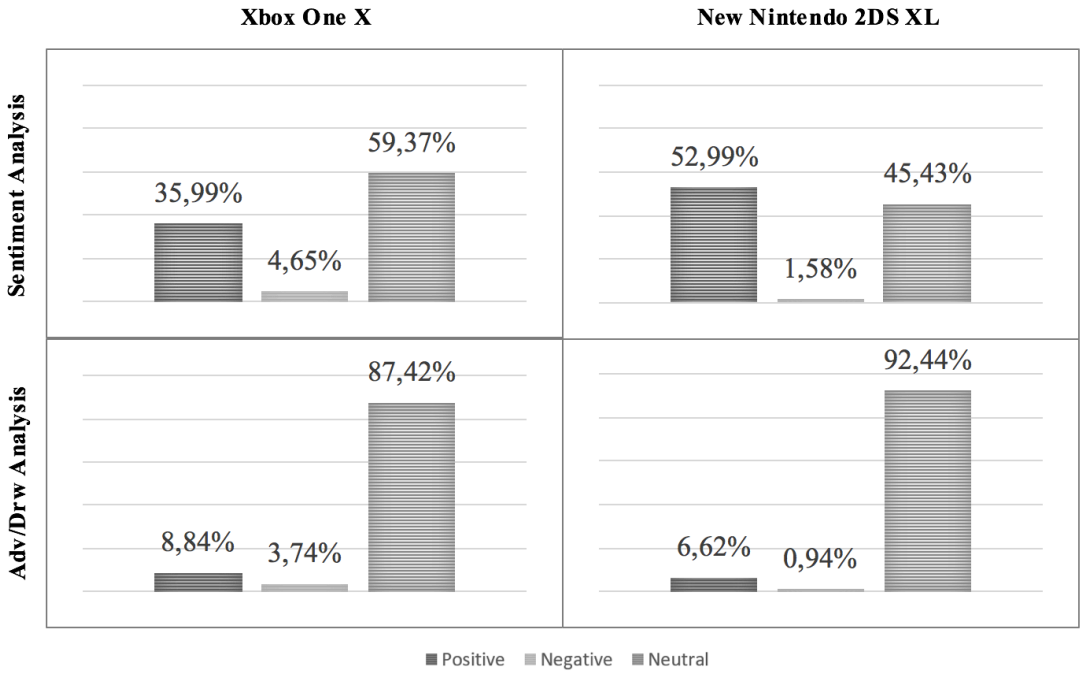

Figure 10.1: Comparison between Sentiment analysis and Advantages/Drawbacks analysis.

found lower levels of neutral tweets, respectively 59.37 % for Xbox One X and 45.43% for the New Nintendo 2DS XL.

This was an expected result since this kind of analysis is designed to deal with colloquial language while our lexicon is technical, being derived from patents analysis. What surprised us is the different polarisation of the products that we see comparing the two analyses. In fact, while with sentiment analysis Nintendo achieves lower percentages of neutral tweets, with advantages and drawbacks analysis is the opposite, since Xbox tweets are more polarised. We also noted that we found more tweets with words of our lexicon in the Xbox subset than in the Nintendo one (10.3). We did the hypothesis that the differences between the percentages of tweets with words found for each product, and the differences of polarisation between the two analyses depend on the different marketing focus, target customer, and technological complexity of the two new video game consoles. Xbox One X targets hard-core gamers who really wants a premium experience [1]. With its marketing campaign, Microsoft pushed the technical supremacy of its new machine over the competitors' products, fuelling a debate about its technical features amongst the potential users.

As a result, the campaign produced a more technical social discourse that allowed us achieving better results. Instead, the new Nintendo handheld console has been developed targeting children and families providing a model that falls somewhere in the middle of the line of 3DS consoles [2]. We initially checked our hypothesis using Google Trend to compare users' search interest about technical review of the two products during the data collection period (Figure 10.2). Then, we analysed the number of technical articles related to the new products published by the 25 most popular video games and technology websites in the U.S, according to the ranking of SimilarWeb, a digital marketing intelligence company which publishes insights about websites. We used Google search engine to retrieve technical article within the web domains previously identified: we obtained 1.117 articles about Xbox and only 52 about Nintendo, proving that technical debate concerning Xbox is greater. This is and evidence of the fact that when a product has a certain technological complexity and fuels a more technical debate, advantages and drawbacks analysis is more able than sentiment in producing technical-functional judgements. The greater number of neutral

[1] http://www.businessinsider.com/why-xbox-one-x-costs-500-2017-6?IR=T (last access: 17/11/2017)

[2] http://www.nintendolife.com/news/2017/05/reggie_explains_the_reasoning_behind_the_new_2ds_xl (last access: 17/11/2017)

Figure 10.2: Google Trends comparison of search-terms "Xbox One X review" and "New Nintendo 2DS XL review" during the data collection period, since 11th June 2017 to 31st July 2017. Values on the vertical axis depict search interest compared to the highest point in the graph during the observation time. A value of 100 is the peak popularity for the term. On average, users searched for Xbox reviews with an approximately five times higher frequency.

tweets found with advantages and drawbacks analysis can also be explained with the Means-end chain model (Reynolds et al., 1995). Consumers express themselves basing on personal consequences linked with product use or basing on personal values satisfied by the product itself. For these reasons, tweets contain a more colloquial language which sentiment analysis is more able to interpret than the latter tool.

Part IV

Applications of the Results

Chapter 11

Exploiting patent information in novel ways

Due to the complexity and volatility of user needs, companies increasingly ask product designers and engineers to create ideas that meet needs in novel and better ways, rather than just making existing technologies more attractive (Brown and Wyatt, 2010). As a matter of fact, these professionals are nowadays involved in the process of understanding in depth what users want and desire (Haley, 1968; Day et al., 1979). Unfortunately, it is well known that user needs are usually examined in separate business departments, such as marketing or business development, and are described in a language that is remote from the professional practice of product designers and engineers. The relation between the understanding of user needs by marketing departments and the development of new products by technical departments is a deeply troubled one. There is a large agreement within the design community that this state of affairs is not optimal and that dedicated efforts should be made to reconcile the engineering approach with a more articulated understanding of user needs, particularly of consumer needs (Pahl and Beitz, 2013; Eppinger and Ulrich, 1995).

A promising approach is based on the description of products in terms of advantages and disadvantages, or drawbacks. Users typically choose an artifact considering the advantages that it brings and the disadvantages that it solves. Advantages and drawbacks exist if they have an impact on the user and if they affect the product in terms of effectiveness (the level at which the product reaches its goals) and efficiency (how many resources does the product have to consume to reach its goals).

At the state of the art, the two main tools to manage advantages and drawbacks developed by the design community are QFD and FMEA/FMECA. Companies frequently make use of Quality Function Deployment (QFD) in order to generate lists of requisites, users' needs, users' requirements and to guide the design process (Carnevalli and Miguel, 2008). They use FMEA/FMECA to gather and study drawbacks, failure modes and their effects and causes (Liu et al., 2013). On the other hand, the notion of advantages is also at the core of marketing techniques used in the segmentation of markets (benefit segmentation) and in the identification of alternative design solutions to achieve desired benefits (means-and-ends-chain analysis).

The interest in the description in terms of advantages and drawbacks is that it can be interpreted smoothly from the two sides of this troubled relationship: engineers can easily link them to performance specifications (usually described with a functional language) and hence technical specifications, while marketing experts can read them with the language of social sciences (for example, psychology, semiotics, sociology or anthropology). Given the promise of this description, why is it used so rarely?

There are several reasons. First, information on user needs is typically owned by users, and is stored in implicit and non codified formats. Second, and consequently, in order to access this information product developers must enter into direct and personal contact with users, listening and understanding the voice of the customer. Not only this is very expensive, but the experience shows that the earlier the stage of development of needs, the more ambiguous, fuzzy and uncertain the information obtained by users. Third, most of this information is not publicly disclosed but is kept confidential as company know-how. Researchers

have hard time to access structured analyses of products based on advantages, even more so for descriptions based on drawbacks. Thus the goal of building up full scale descriptions based on advantages and drawbacks is still elusive.

In the present work we consider patents as a possible alternative information source for advantages and drawbacks. As stated by the World Intellectual Property Organization (WIPO), an invention is a solution to a specific technological problem (Organization, 2004). The problem that an invention solves in a technological field is a certain negative effect that the state-of-the-art technologies cannot overcome; on the other side, a solution is the way to solve this problem. A solution can lead to some advantages with respect to the known state of the art. Thus, starting from the definition of invention, it is clear how it can be characterized by its advantages and the problems it solves. Based on these definitions, the WIPO explicitly suggests as a guideline for applicants to write patents in this language. The applicant (the person or company that applies for the patent) is led to include this information in patent documents in order to have more chances of success in the patenting phase.

An important feature that makes patent information valuable is that the information that is contained in these documents today will be contained in other documents, like manuals, handbooks and market reports, to which designers are more accustomed, in the future: information anticipates availability of products on the market by a factor varying from 6 to 18 months (Golzio, 2012). In addition, these documents are freely accessible by many different databases nowadays (Kim and Lee, 2015).

To claim that patents include descriptions in terms of advantages and drawbacks is one thing, to show how this information can be used effectively, however, is a completely different business. To test the hypothesis of the presence of advantages and drawbacks information in patents and to exploit this information for design purposed, there is a need to overcome two main problems:

- analyzing patents requires skilled personnel and long time (León-Rovira and Cho, 2007)
- due to the increase in the number of patent publications, there is a massive information overflow (Bergmann et al., 2008).

In this chapter we present a methodology for the extraction of information on advantages and drawbacks of technologies from patents, that is able to fully overcome these problems (as demonstrated in section 7.2) and with the final goal to make available patent-based structured information to the design community.

11.1 Towards formal definitions of Advantages and Drawbacks

Referring to section 7.2, we propose that all useful definitions of advantages and drawbacks can be collapsed into three categories, each with a positive or negative sign, as follows:

1. more/less wanted output obtained . A wanted output is a desired effect of the system.
2. more/less unwanted output obtained. An unwanted output is undesired effect of the system.
3. more/less resources needed. More resources needed to achieve a desired effect imply less efficiency.

This classification can be labelled ADIO classification (Advantages-Disadvantages-Input.Output). The operationalization of this classification for purposes of automatic information extraction and processing is the object of the rest of the paper.

11.2 Methodology

The goal of our system is to automatically extract short sentences that contain information about the advantages and the drawbacks of the technology from patent texts. Furthermore we propose a taxonomy that organizes the output of the system focusing on advantages and drawbacks that have impacts on the systems thus influencing its input or output. A flow diagram representing the adopted method for the automatic ADIO extraction and classification of a technology is shown in Figure 11.1. The method takes as entry a patent set representing the technology to analyze. The patent set and the list of advantages and

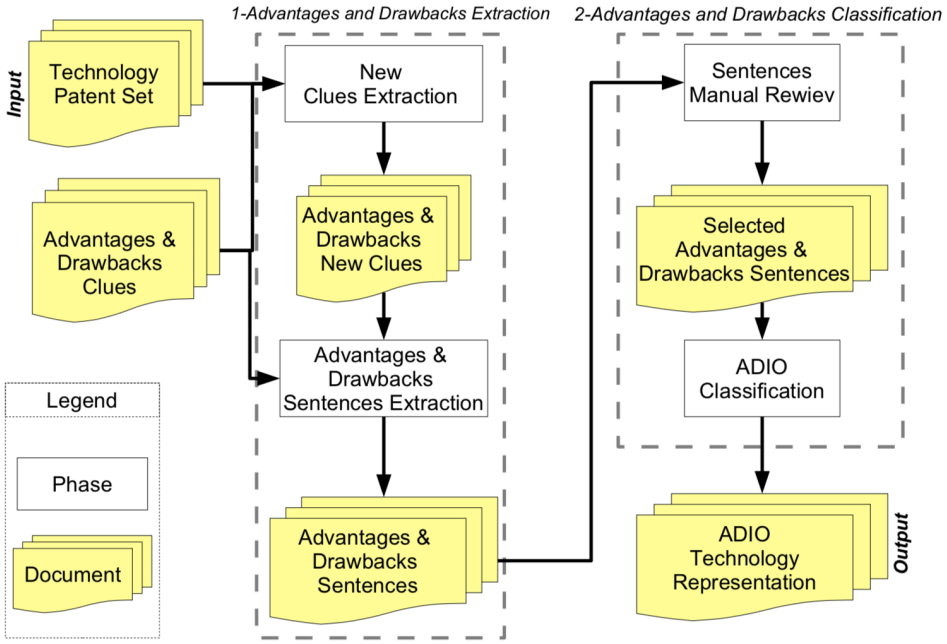

Figure 11.1: Work flow diagram followed to extract the ADIO technology repres- entation from a patent set.

drawbacks clues are entries of the process of advantages and drawbacks extraction and generate the phrases containing the advantages and the drawbacks. Than they become the entry of the process of Advantages and Drawbacks classification that exploits human knowledge to classify the technology according to the ADIO representation.

11.2.1 Advantages and drawbacks extraction

The process of advantages and drawbacks extraction is the first of the two-macro processes used in our system. The first process starts from a patent set containing patents inherent to a technology and extract relevant sentences in output. Each sentence describes an advantage or a drawback of the specific technology. All steps of this process are fully automatic. The patent set should be very large, in the order of several hundred (in our case study n $>$ 1000 items). To describe with a certain degree of precision an advantage or a drawback, patent applicants have to use sentences of a certain length. Since NER systems are designed to extract single words or short n- grams, we need to extract entities that are clues of the whole sentence that describes the advantage or the drawback. However our interest is not on the clue but rather on the words that follows the clue: the real advantage or the real drawback. We refer to these words as target. Considering the ADIO classification, proposed in the present work, these are words that help to classify whether the advantage or the drawback have an impact on the input (influencing efficiency) or the output (influencing effectiveness) of the system. The few examples above shows how clues are words that indicates a characteristic of a flow or its modification (positive or negative); the clue and target together specify the entity and direction of the modification of the flows that evolve within the system. A summary of these linguistic concepts and some examples are shown in figure 11.2.

As stated above, we introduce a crucial concept, that of a “clue” for the identification of a complex text structure describing advantages or drawbacks. We describe here the process for collecting clues. The process is not trivial because the sources are heterogeneous, fragmented and sparse. For example, we can find lists of failures in repositories published in the car industry, lists of diseases or infections in medical treatises, lists of positive and negative words in sentiment analysis tool-chains. Some of them are very accurate but extremely short, domain specific and rarely occurring in patents (e.g. new diseases), while others are broad but ambiguous thus introducing noise in the analysis (e.g. sentiment annotated words). We followed a

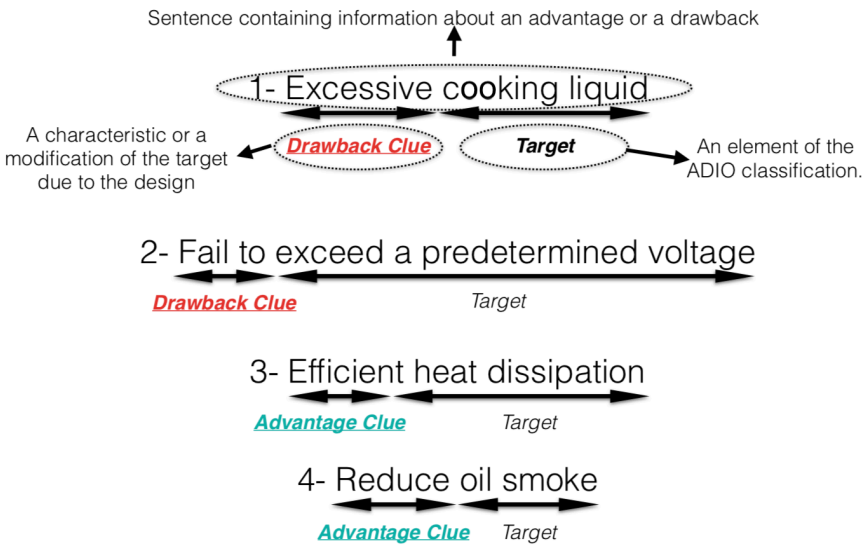

Figure 11.2: Examples of advantage or drawback sentences divided in its clue and target.

twofold approach. The first approach consisted in the manual collection of clues of advantages and drawbacks directly from patent texts. This process was performed on 2000 patents in several patent classes. This has led to collect 3.254 advantage-clues and 5.142 drawback-clues. The second approach consisted in looking for alternative methods to indicate advantages or drawbacks clues, finding defined word patterns. The most relevant are the negations of advantages to obtain drawbacks, and the negation of drawbacks to obtain advantages. It is worth noting the cases of suffixes like as -less or -friendly, - free and the like, and prefixes like as anti-, dis- , de-, un- and the like, that allow a rapid and systematic expansion of the database. At the end of the process, advantages numbered 6.568 and the drawbacks numbered 14.809. This is a fairly large knowledge base for the system, and gave us a reasonable number of clues to be used in the next step of the process. Example of clues are shown in section 7.2. The first approach has the limitation that lists were extracted from a random but relatively small sample of patents (n= 2000). Another limitation is that the rules used in the second approach are not exhaustive, and they can create non-sense clues, due to the possible combinations of words (e.g. “anti- ability” or “un-problem”). On the positive side, it is reasonable to assume that using these approaches it is possible to collect a large set of clues that are relatively independent from the patent set. In addition, it is now clear how new clues could be easily extracted when changing patent sets. In order to obtain a larger and complete collection of clues it is unsuitable to use the manual extraction on each domain patent set. For this reason, new clues were iteratively used to train machine learning algorithms.

New Clues Extraction

In this section we briefly describe the system used to automatically extract new word clues from patent texts. The system is based on the work discussed in section 7.2. This process takes in input a corpus of patent documents regarding a certain technology. After the tokenization of the corpus, each token (word or n-gram) is represented by series of features. Then the advantage and drawback clues are re-projected on the text, generating a training set of words to be given as an input to a classifier system. The classifier builds a model able to detect words that have similar behavior (in terms of the selected features) with respect to the behavior described in the training set. The model is used to classify the words contained in patents as potential new advantages or drawbacks word clues. These new words clues are technology specific clues or generic clues that did not belong to the starting list of advantages and drawbacks generic word clues.

Advantages and Drawbacks Phrases extraction

Once all the new advantages and drawbacks clues are extracted, these are merged with the ones belonging to the original knowledge base, obtaining a final list which will be processed by the advantages and drawbacks sentences extractor. The advantages and drawbacks sentences extraction is the activity through which the system catches the shortest informative sentence containing each word clues. To do that the patents are processed through a phase of part-of-speech tagging (POS tagging). Starting from the clues, only the POS sequences that match a certain pattern were extracted. The pattern, expressed using a regex regular expression is:

$$(Clue) + Noun. * Noun. * Noun. *$$

This structure has proven to be able to catch a reasonable number of words of the target, exhaustively expressing an advantage or a drawback without catching very long phrases.

11.2.2 Advantages and Drawbacks ADIO classification

The process of advantages and drawbacks classification is the second of the two macro processes involved in our system. This process takes in input the advantages and drawback sentences extracted in the advantages and drawback extraction process and gives in output the ADIO representation of the technology.

Sentences Selection

As stated above, we suggest a clear classification of advantages and drawbacks in a 3*2 structure. After the extraction each sentence is assigned to one of the following classes:

1. more/less wanted output obtained. A wanted output is a desired effect of the system.
2. more/less unwanted output obtained. An unwanted output is undesired effect of the system.
3. more/less resources needed.

If a sentence does not belong to one of these classes it is not taken in to consideration for the next analysis, even if expresses advantages or the drawbacks of the invention. This classification makes it possible to represent the technology using the ADIO representation.

ADIO Technology Representation

Given the classification described above, we obtain three possible kind of advantages and three possible kinds of drawbacks. Considering a wanted or desired output, the achievement or the increase is an advantage, while the negation or the reduction is a drawback. On the other side, considering an input to the system or an unwanted output, negation and reduction constitute an advantage, while achievement and increase are clearly a drawback. It is important to specify that the both the input or the output (wanted or unwanted) could involve flows of matter, energy, or signal.

11.3 Results

11.3.1 Patent set

To test the proposed process, we selected a patent set composed of a sample of 3,000 patents. The patent sets belongs to the A47J37 IPC patent class defined as *"Baking; Roasting; Grilling; Frying"*. We will refer to this patent set as cookers set.

11.3.2 Extraction of Advantages and Drawbacks

Total extracted advantages numbered 4129, drawbacks numbered 1835. After manual review of sentences the total number went to 2509 and 1532, respectively. During the manual review phase each sentence was assigned to one of the three classes of the taxonomy, considering the target of the sentence. The system itself decides if a sentence indicates an advantage or a drawback, considering the clue. The results in terms of cardinality of the classes of the taxonomy are shown in Table 11.1. As we can see from this table, the sentences review process has led to a balance between the extracted advantages and drawbacks. It is interesting to see how the wanted outputs are more likely expressed as advantages (1786 sentences are advantageous wanted output while 660 are drawbacks); the situation is reversed for the unwanted output (431 sentences or that advantages and 682 for the drawbacks).

Table 11.1: Examples of clues of advantages and drawbacks extracted manually from patents.

Class	Number of Advantages Sentences	Number of Drawback Sentences
Input	292	190
Wanted Output	1786	660
Unwanted Output	431	682
TOT	2509	1532

11.3.3 A-D-I-O Representation

The two ADIO schemes for the advantages and drawbacks of cookers are shown respectively in Figure 11.3. The sentences shown in this figure are a sample of all of the 2662 extracted sentences. Furthermore these sentences are taken as-is from patents, misprints and errors included.

Both the results are promising for future applications in the design fields. In particular Figure 11.3 (a) allows designers to focus on the positive side of the effects provided by the product and to better meet the explicit and implicit user needs. Similarly, Figure 11.3 (b) helps designers to redesign of the product in a proactive way, to keep attention to the critical issues identified by the drawbacks and to conceive possible corrective actions to solve such drawbacks.

11.4 Discussion

This paper has proposed a method to extract and summarize sentences that describe advantages and drawbacks of technologies from patents. Advantages and drawbacks are considered as phenomena that influence the efficiency or the effectiveness of products by modifying their inputs or their outputs. Advantages and drawbacks information are useful for designers who want to design new products or to redesign old ones so to meet user needs in novel and better ways. The proposed approach allows patent readers to analyze a massive quantity of patents and to reduce the time needed for research and analysis.

In the future, we want to focus on the application of the proposed ADIO framework to a wider number of patents set and hopefully we would like to automate the classification of advantages and drawbacks. Furthermore we want to focus on the extraction of new entities of interest for designers, and to understand which other groups of words contained in patents texts can add value to the design process.

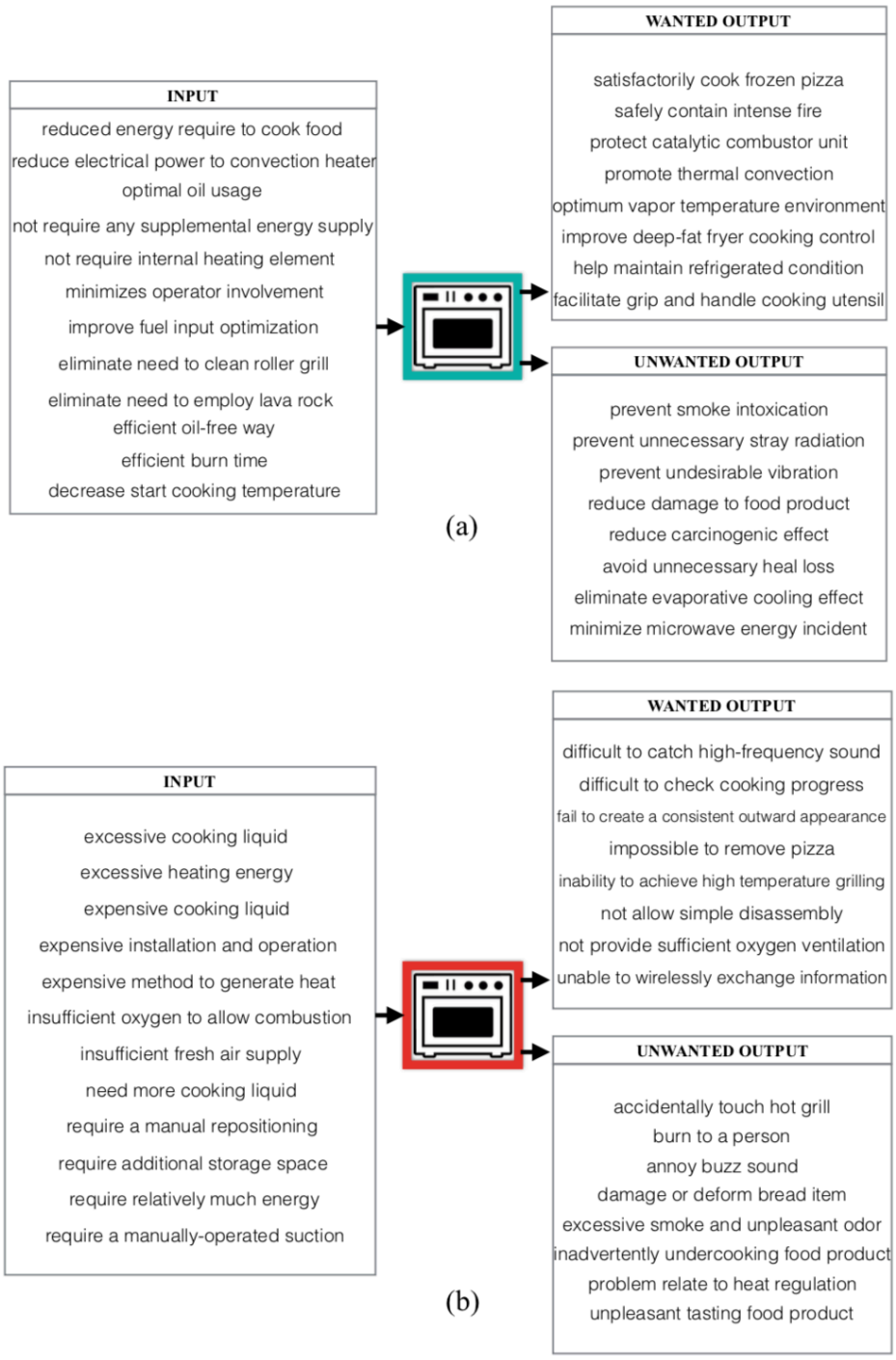

Figure 11.3: Examples of advantage or drawback sentences divided in its clue and target.

Chapter 12

Enriched dictionaries for Innovation

The demand for intelligence and foresight of technologies is increasing due to the need of companies and governments to make sense of the rapidly changing technology landscape and to make better decisions. In particular, emerging technologies exhibit not only rapid growth, but also strong conditions of technology and market uncertainty, so that traditional techniques of technology intelligence are challenged (Rotolo et al., 2015).

Technology intelligence makes large use of a statistical procedure called clustering. This multivariate technique is commonly used to place entities into relatively homogeneous groups, maximizing the difference with other groups, when entities are not subject to an existing classification. In technology intelligence this is the most interesting situation: if technologies were already fully classified, then they would already be established or mature.

It is important to remark that clustering, in one way or another, is almost a necessity in technology intelligence. The amount of data available on technologies, even on the last generation of technologies, is so large that a preliminary effort of clustering and profiling is generally considered a preliminary step for the analysis.

In technology intelligence the formation of clusters is generally based on words extracted from relevant documents (scientific publications, patents, technical standards). There are two main approaches to the extraction of words from documents: extracting from the metadata associated to the document, or extracting from the full text of the document. Examples of metadata are authors, affiliations, inventors, assignees, keywords, titles. This approach has generated a large literature that uses metadata in order to extract usable knowledge. Within this literature, a notable stream of studies has extracted usable knowledge from documents by clustering the metadata in order to obtain meaningful structures. This method is associated to the use of keywords, as we will see below.

More recently, a different approach has been introduced, based on the processing of the full text of documents. Following the remarkable advancements in computational linguistics, it has become possible to process the entire text of publications, patents, or technical standards to a large scale. Here the clustering exercise does not take place on metadata, but on words, or their combinations, included in the text of documents. The topic modeling is the most used approach. Topic modeling is a tool to extract structures of text from corpora without the help of external sources of knowledge (Blei et al., 2003). Once the words have been extracted from metadata or the full text, a clustering strategy must be designed based on an appropriate definition of similarity.

We then suggest the exploration of a novel approach, one that combines domain knowledge with powerful data science techniques, called *enriched dictionaries*. These are large and highly structured collections of words, associated to formal definitions and internal linkages, that are produced on the basis of domain knowledge of various kinds. In some cases they are publicly available, in othr cases they are the resuld of dedicated and idiosincratic research efforts. These dictionaries can be used in order to "filter" the semantic content of the full text of documents according to pre-defined structures, generated within the domain

knowledge and validated at the state of the art. They can be used, therefore, within the so called supervised text mining approach. This paper has two objectives: introducing the methodology of enriched dictionaries, and showing that it allows the joint use of several, and complementary, perspectives: one based on the abstract engineering principles of technologies (functional view), another on the advantages delivered by the technology (advantage view). These dimensions are kept separate in the literature and clustering exercises do not combine them. We show the power of clustering technologies using these views in a combined way.

12.1 An overview of dictionaries for technology intelligence

12.1.1 Publicly available dictionaries

One of the advantages of the dictionary approach is the possibility to utilize publicly available dictionaries, that is, available online with unrestricted, free access. We made systematic use of Wordnet, a large dictionary in English language made available and curated by Princeton University. Wordnet is a standard reference in computational linguistics.

Upon the basic Wordnet dictionary, several specialized or dedicated dictionaries have then been developed, which are also available with unrestricted, free access. We made use of the following dictionaries.

- Artifacts: a collection of terms referring to objects, inventions and tools. The dictionary is one of the Name categories available in the Wordnet-based categorization dictionary from Provalis, free to download [1]. Artifacts were chosen to explore texts in order to find structural indications for the components of an invention.
- Acts: a collection of terms referring to generic actions. This is another Name category from the Wordnet dictionary. Acts were selected in order to find verbal expressions referable to functions carried out by the objects described therein.

12.1.2 Research-based dictionaries

The dictionary methodology offers the flexibility to build up dedicated or specialized dictionaries, that represent the state of the art of a given knowledge domain. We make use of two dedicated dictionaries recently developed by the authors in an academic research context. Needless to say, in these cases the authors must give the readers or users of the dictionary full proof of the completeness of the dictionary, or the compliance with formal criteria for the definition of a dictionary. These proofs have been offered in a number of papers cited below.

- Functional verbs: list of verbal expressions describing functions, or actions performed by artifacts on given objects (Fantoni et al., 2013; Apreda et al., 2016).
- Advantages and Disadvantages: list of terms referring to measurable benefits or drawbacks related to an object, chosen for their capacity to represent applications of an invention (see section 7.2).

Table 12.1 offers an overview of the dictionaries used for this analysis, with a few examples of terms.

Table 12.1: Sources and main characteristics of dictionaries.

Dictionary	Numberof entries	Examples of chunks
Artifacts	12374	Acetanilid, Actuator, Aerogenerator, Agglomerator
Acts	5527	Abduction, Abidance, Accenting, Acquiring
Functional Verbs	11256	Abrade, Absorb, Abut, Accelerate
Advantages/Disadvantages	29902	Ability, Accelerate, Acceptance, Accessibility, Accident, Accidents, Aggravated, Aggravation

[1] https://provalisresearch.com/products/content-analysis-software/wordstat-dictionary/wordnet-based-categorization-dictionary

Dictionary	Numberof entries	Examples of chunks
Artifacts	12374	Acetanilid, Actuator, Aerogenerator, Agglomerator
Acts	5527	Abduction, Abidance, Accenting, Acquiring
Functional Verbs	11256	Abrade, Absorb, Abut, Accelerate
Advantages/Disadvantages	29902	Ability, Accelerate, Acceptance, Accessibility, Accident, Accidents, Aggra

12.2 The value added of enriched dictionaries

In order to illustrate the original contribution of the enriched dictionary approach, it is useful to refer to similar approaches, recently introduced in the literature. The starting point of this literature is similar to the one advocated in this paper, i.e. the need to supervise the full text mining with the help of list of words that reflect domain knowledge.

A natural candidate here is the list of words that describe technical functions, or the abstract characterization of the working principles of technologies (Cascini et al., 2004; Dewulf, 2006; Cascini et al., 2007; Cascini and Zini, 2008). The pioneering approach in this field is TRIZ, the patent-based methodology that has identified a number of abstract inventive principles (Petrov, 2002), whose application to patent texts has permitted the detection of evolutionary trends in specific technologies (Verhaegen et al., 2009; Wang et al., 2010; Yoon and Kim, 2011a; Park et al., 2013b). A variant of this approach combines the functional approach with lists of product attributes or properties (Yoon and Kim, 2012; Yoon et al., 2011; Kim et al., 2010). More recently, a similar approach has been introduced, suggesting that subject-action-object (SAO) linguistic patterns can be derived automatically from the full text of patents (Yoon and Kim, 2011b; Choi et al., 2012; Park et al., 2011b). A SAO textual sequence is considered an indication of the engineering principle that describes an action that is changing an object. Choi et al (Choi et al., 2012) reviews the state of the art of function-based technology databases (in particular, TRIZ and Creax) but suggest that SAO structures are more versatile and flexible in order to apply Natural Language Processing techniques.

Within this line of investigation the notion of technology tree (Tech Tree) has been introduced (Choi et al., 2012). It combines within a single representation taxonomies of products, technologies, and abstract functions. Similarity matrices are built along these dimensions and aggregated.

While these two approaches have generated a large and interesting literature, their main limitation is the lack of transparency. The queries obtained from the TRIZ inventive principles have not been published and there is no demonstration of their reliability (i.e. replicability under controlled conditions) with respect to all semantic variations of words. Similarly, the SAO queries, although intuitively clear, leave the readers and users with the problem of establishing the semantic content in the engineering sense, given that the notion of "action" may cover, in fact, many different meanings.

12.3 Methodology

The cited studies have suggested that a mixed approach to the clustering problem could be the best solution to achieve good results. In particular there is a need for placing more domain knowledge into the analysis, while keeping the enormous advantages of text mining and automatic knowledge representation techniques.

This is the starting point of the dictionary approach we propose. It suggests that the full text of technical documents is searched by using a structured list of words, generated from a substantive body of domain knowledge, formally defined, and organized in a hierarchical way.

A dictionary, in substance, is a collection of terms pertaining to a precise context and then selected following a logical pattern. The content of a dictionary may sensibly vary from case to case, comprising terms related

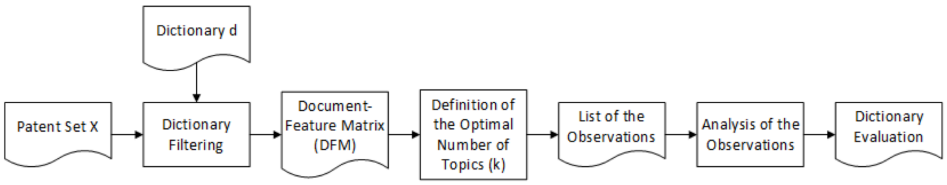

Figure 12.1: Steps of the dictionary approach to the text mining of patents.

to a precise technology, a technical field, a social group, or more generic parts-of-speech such as particular kinds of actions or nouns, and so on.

The main difference between a dictionary and a mere collection of keywords is in their respective scope, which is, for a dictionary, wider and far more complete, including all the synonyms, hypernyms and hyponyms of a term. By definition, a dictionary must include all the definitions and terms referring to the chosen context. The completeness has an important consequence for text mining applications. When a dictionary is used as a tool to filter the content in a collection of texts, the query is formed not only by principal words, but also by secondary words, such as synonyms. These secondary words would be lost if the analysis were based on words and keywords. Therefore dictionaries contribute to overcoming the biases introduced in the analysis by the subjective choice of keywords by experts.

To start with, the dictionary method was combined with topic modelling techniques in order to overcome one of the main limitations of topic modeling, that is, the extraction of irrelevant topics.

The starting assumption is the possibility of representing a document using the main topics cited in it. A topic is, basically, a collection of terms that delineate and describe an argument. Topics may be identified by using unsupervised or supervised methods. The unsupervised application of the method will have the result of giving generic topics, often irrelevant in specialized analysis. This outcome can be mitigated, but not eliminated, by the application of ranking techniques, such as Term Frequency-Inverse Document Frequency (TF-IDF), that allow the identification of coherent topics. Alternatively, analysts may refer to supervised methods, which are however resource-consuming in their design and implementation. The dictionary approach offers an excellent solution to the trade-off between relevance and precision of the search, which plays in favor of supervised methods, and the parsimony of the analysis, which on the contrary militates for the adoption of unsupervised methods.

In addition, if topics are clearly delineated, it becomes possible to identify similarities between documents that refer to technologies with a transversal, or cross-domain, potential for application. Dictionaries can be used for different purposes and help to identify latent structures from a variety of perspectives. We advocate the use of several dictionaries in parallel, on the same collection of texts, as a way to examine technologies with a variety of perspectives. Figure 12.1 illustrates the main steps of the methodology.

The basic idea of the process is that a document can be represented by the set of terms, numbers, graphic symbols and punctuation that compose meaningful sentences. The “technical” objects belonging to this set are named Features. Starting from selected patent sets, a Document-Feature Matrix is created using different dictionaries as filters to select only the desired features; the obtained matrices are then used as the starting point to perform a topics number evaluation algorithm, based on the connections between documents with same or similar features. Topics are formally clusters of features, coupled because of their pertinence to the same concept; measuring the percentage of cohesion of a document to each topic recognized is, then, a method to collect informations about the contents of the document itself.

Creation of the Document-Feature Matrix

A patent can be represented by a vector exhibiting the number of occurrences, if a given feature is included, and 0 otherwise. Since these features are ordered within dictionaries, all patents in a patent set can be represented with similar vectors. A collection of texts can be therefore represented as a matrix called

Document-Feature. A Document Feature Matrix (DFM) is a algebraic matrix with N rows, corresponding to individual documents in the corpus, and M columns representing the features. With this formalization it is possible to measure the occurrences of features in every document contained in the corpus. It is a powerful tool to get a fast quantitative information about a document set, counting the features originated in the dictionaries, with the opportunity of selecting (or removing) some of them.

Each patent in the patent set is filtered using one of the four dictionaries at a time. The resulting list of words was processed in order to eliminate the generic words, or those words that, being diffused across most documents, do not have specific semantic content. This is done by using the stopwords labelled SMART in the text mining literature, in combination with a list of stopwords extracted from the text of patents and including some generic words in the legal language of patents (or "Patentese"), such as claim, invention, right, tool etc.

Due to the large number of entries in dictionaries and the exploratory nature of this work, we restricted the number of features considered for each of the DFMs to 500. The selection was made automatically, by selecting the top 500 terms in terms of share of documents in which they appear. In some cases the selected patent text delivered no matching with the selected features. In this case the patents were eliminated. Thus the fina DFM only includes documents with positive entries from the lists of features extracted from the analysis. It must be noted that DFMs are sparse matrices: sparsity is a mathematical parameter, varying from 0 to 1, that indicates how much systems are loosely coupled. In particular, it is a measure of the Z zero-valued elements in a matrix divided by the N x M total number of elements. Thus, a sparse matrix is a matrix in which large part of elements in cells are zeroes. Sparsity it is a good measure of how documents in a corpus differ from each other with respect to their Features.

Documents are iteratively examined in pairs: if two documents share a feature they are considered similar. Various metrics of similarity can be defined and computed. After a measure of similarity is defined, various clustering techniques can be applied as well. In this application we use the cosine similarity, which is largely adopted in the field. Cosine similarity is a measure of similarity between two non-zero vectors, that computes the cosine of the angle between them in an inner product space. Each document in the Document-feature Matrix is characterized by a vector where the value of each dimension corresponds to the number of times the feature appears in the document. Two documents, then, are similar if their cosine similarity is near to 1.

Given two n-dimensional vectors of attributes, x and y, their cosine similarity is calculated by the dot product of them, normalized by the product of the vector lengths. Results of the similarity measures are stored in a N x N matrix, in which documents are both upon rows and columns. It is a squared matrix with ones on its diagonal, indicating identities between patents and themselves, and the upper (or lower, for it is the same) half filled with similarity values for all the document couples. The most intuitive method to analyze which patents are more similar is to visualize them in a graph. To do so, it was used the igraph package , which contains a list of commands for creating and analyzing graphs. We can use our similarity matrices as a base to build the corresponding graphs, using them as they were adjacency matrices.

Adjacency Matrix

An Adjacency Matrix is a square matrix that represents a finite graph. In this case, the similarity matrix has in position (i,j) the inverse of the distance between vertices vi and vj . This gives to the graph not only information about whether two vertices are connected, but also a measure of their connection. It was chosen to set a threshold t = 0.8 for similarities in order to avoid representation of loose relations between vertices. To better visualize results, graphs were migrated from R to Gephi, a visualization software that simplifies this kind of operations. Gephi makes possible to assign colours to nodes as an intuitive way to label them; in our graphs four different colours were used to partition patents in their class, then Force Atlas layout algorithm was performed to see how well the graphs were clustered.

Optimal Number of Topics

Once created, a DFM can be interpreted as a graph showing similarity links between documents. In this context, in fact, similarity can be defined in terms of the percentage of total features that are shared between documents. Features can also be grouped in topics. In turn, similarity between documents can be defined in terms of the percentage of topics shared between documents. Finally, documents that refer to the same topic, hence are similar among them, can be grouped in cluster. Within a cluster it is possible to define a metrics of cohesiveness: the larger the number of topics shared by documents, the larger the cohesiveness of the cluster they belong to.

The optimal number of topics is not defined once and for all, but it depends in subtle ways from the characteristics of the documents in the set and the list of features. Thus the optimal number is the result of an experimental approach. The R package ldatuning calculates several metrics and then delivers an estimate on the optimal number of topics for Latent Dirichlet Allocation (LDA) models. In our application, we chose the CAOJUAN metric to evaluate the number of topics. The decision to use just one metric was made due to computational reasons; we selected the model proposed by Cao Juan, since he demonstrated that LDA models perform in an optimal way when the average cosine distance of topics reaches the minimum. Starting from the assumption that less correlated topics are more independent, the author defined average dispersion as a metric to measure the topic structure, using cosine distance. A smaller average dispersion denotes a more stable topic structure, that is, a topic model which is defined in a better way.

12.4 Results

We now turn to the analysis of the topics resulting from the filtering procedure based on dictionaries, the construction of the DFM, the measurement of cosine similarity. We will use four measures to highlight the content of the patent sets:

- Average cosine distance of topics
- Average number of topics
- Average sparsity of DFM

The dictionaries were divided in two macro-classes, due to their nature: Non-Technical Dictionaries (Acts and Artifacts) versus Technical Dictionaries (Functional Verbs and Advantages & Disadvantages). The Blacklist set, filtered with canonical and "Patentese" stopwords, will be useful for a more complete understanding of the findings.

In order to test the dictionary approach to clustering, we trained the algorithm on a standard patent set, formed by patents belonging to four IPC classes whose technologies are largely known. Table 12.4 illustrates the International Patent Classification (IPC) classes included in the analysis (a random sample of 2.000 patents each).

Table: Patent sets under examination .

IPC Class	IPC Definition
A47J37	KITCHEN EQUIPMENT; BAKING DEVICES; ROASTING DEVICES; GRILLING DEVICES; FRYING D
A61C15	DENTISTRY; APPARATUS OR METHODS FOR ORAL OR DENTAL HYGIENE
A61G13	TRANSPORT, PERSONAL CONVEYANCES, OR ACCOMMODATION SPECIALLY ADAPTED FOR PA
A61H	PHYSICAL THERAPY APPARATUS, e.g. DEVICES FOR LOCATING OR STIMULATING REFLEX PO

The technologies examined are largely mature. As it is shown in Figure 12.2, in some of them the patenting activity started as early as XIX century. The maturity of these technologies is a good starting condition for the training set: similarities and dissimilarities of the patents should be clearly visible.

By placing them together in the same patent set we deliberately include very different technologies. We ask the algorithm based on dictionaries to be able to discriminate technologies in a patent set which is, by construction, highly heterogeneous.

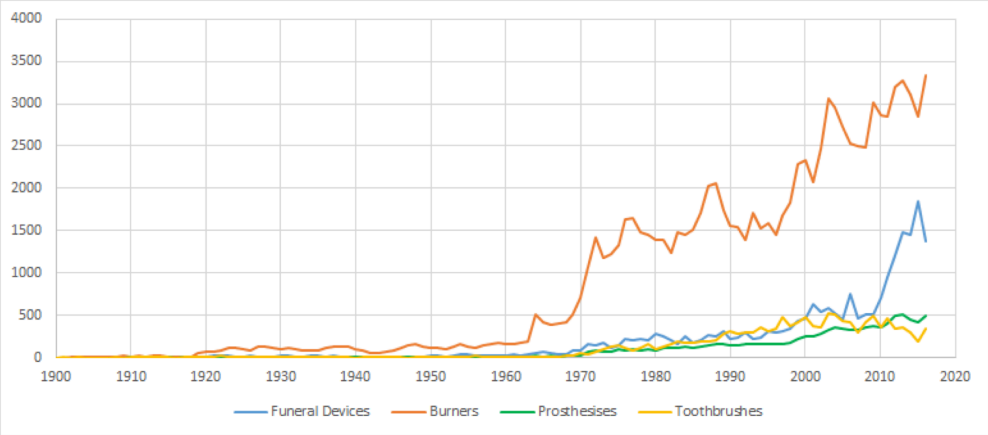

Figure 12.2: Technological history of the technologies in the selected patent sets.

Table 12.4 offers a summary of measures derived from the application of dictionaries to the four patent sets combined together.

Table: Summary of findings from the application of dictionaries to the combined patent sets.

Dictionary	Average cosine distance of topics	Average number of topics	Average sparsity of the D
Artifacts	0.002	16	9
Acts	0.004	11	9
Advantages and Disadvantages	0.016	32	9
Functional Verbs	0.029	37	7
Blacklist	0.031	40	6

Two findings are visible in Table 12.4. First, there is large difference in the cohesiveness of the clustering based on Technical and Non-technical dictionaries. The average cosine distance calculated with the Artifacts and Acts dictionaries is ten times smaller than the average cosine distance for Advantages & disadvantages and the Functional dictionary. Second, Non-technical dictionaries give origin to a much smaller number of topics.

Let us examine the differences between dictionaries more in detail. We start by examining the performance of each dictionary in extracting and clustering topics.

Non-technical Dictionaries

Table 12.4 shows the results of the application of the Acts dictionary to the four separate patent sets, while Table 12.4 shows the same for the Artifacts dictionary.

Table: Findings from the application of the **Acts** dictionary to selected patent sets.

Dataset	Number of topics	Cosine distance of topics	Number of documents	Number of features	Average spa
Burners	2	0.001	496	306	
Toothbrushes	6	0.002	494	311	
Prosthesis	18	0.005	491	493	
Funeral Dev.	16	0.007	491	470	
Mean	11	0.004	493	395	

Table: Findings from the application of the **Artifacts** dictionary to selected patent sets.

Dataset	Number of topics	Cosine distance of topics	Number of documents	Number of features	Average spa
Burners	6	0.001	495	397	
Toothbrushes	16	0.001	469	500	
Prosthesis	18	0.003	487	500	
Funeral Dev.	22	0.002	485	500	
Mean	16	0.002	484	474	

These two dictionaries were retrieved from the Wordnet categorization dictionary, which is a tool for linguistic analysis; thus, the terms comprised in them are of common use, and their suitability for the description of patents is not guaranteed. Taken together the Non-technical dictionaries deliver two main results. First, there is a difference in the optimal number of topics between simple technologies (Burners and Toothbrushes) and more complex and articulated technologies (Prosthesis and Funeral devices). In the case of the Acts dictionary, the optimal number of topics is 2-6 for the former technologies and 18-16 for the latter. Second, Non-technical dictionaries deliver clusters with an extremely low cosine distance, that is, with strong internal cohesiveness. In other words, Non-technical dictionaries produce an excellent clustering of documents. This suggests that they can be used to obtain a small number of non-technical topics, that allow a general identification of the content of documents, but not the identification of technical similarities and differences.

To provide an example of the potential of the dictionary methodology we show in Tables 12.4 and 12.4 the top ten words found in a selection of the topics extracted with the Non-technical dictionaries in the patent set Prosthesis. We need to test whether Non-technical dictionaries deliver topics that are satisfactory with respect to their potential to describe technologies.

Table: Top ten words in selected topics extracted in the patent set **Prosthesis** with the aid of the **Acts** dictionary.

Topic 3: Term	Topic 3: Beta	Topic 7: Term	Topic 7: Beta
Measuring	0.322	Manipulation	0.160
Measurement	0.217	Steering	0.152
Detecting	0.210	Dialysis	0.145
Performing	0.033	Cycling	0.063
Sending	0.031	Deactivation	0.060
Calibration	0.030	Dampening	0.049
Supplying	0.024	Shearing	0.047
Repositioning	0.024	Logging	0.032
Causing	0.009	Exercising	0.029
Comparing	0.009	Reaching	0.023

Table: Top ten words in selected topics extracted in the patent set **Prosthesis** with the aid of the **Artifacts** dictionary.

Topic 5: Term	Topic 5: Beta	Topic 14: Term	Topic 14: Beta
Backrest	0.498	Electrode	0.405
Eyewash	0.092	Catheter	0.233
Armrest	0.085	Circuitry	0.106
Gasket	0.060	Sensor	0.026
Capacitor	0.052	Housing	0.025
Arms	0.048	Converter	0.025
Gurney	0.045	Emitter	0.021
Footplate	0.025	Cathode	0.018
Eyepiece	0.019	Anode	0.018
Headrest	0.011	Reflector	0.011

In the case of the Artifacts dictionary, Topic 5 gives us information on components related to the body part interested by the prosthesis, while Topic 14 tells us about electrical tools and sensors that are involved in the invention. Also in this case, the two groups are useful to gather information about the basic composition of the set, but not enough on focus to have a satisfactory definition of the technologies.

Technical Dictionaries

Table 12.4 shows the results of the application of the Advantages & Disadvantages dictionary to the four separate patent sets, while Table 12.4 shows the same for the Functional dictionary.

Table: Findings from the application of the **Advantages and Disadvantages** dictionary to selected patent sets.

Dataset	Number of topics	Cosine distance of topics	Number of documents	Number of features	Average spa
Burners	32	0.015	500	500	
Toothbrushes	30	0.014	500	500	
Prosthesis	32	0.021	497	500	
Funeral Dev.	34	0.014	495	500	
Mean	32	0.016	498	500	

Table: Findings from the application of the **Functional** dictionary to selected patent sets.

Dataset	Number of topics	Cosine distance of topics	Number of documents	Number of features	Average spa
Burners	38	0.021	500	500	
Toothbrushes	36	0.022	500	500	
Prosthesis	34	0.032	497	500	
Funeral Dev.	38	0.041	495	500	
Mean	37	0.029	498	500	

The technical dictionaries applied contain a large number of features, generating a large DFM. An overall look at Tables 12.4 and 12.4 shows several interesting findings. First of all, the technical dictionaries allow the identification of a much larger number of topics with respect to non-technical dictionaries: the average number is 32 for the Advantages & Disadvantages, and 37 for the Functional dictionary. This means that technical dictionaries show a higher power of discrimination of semantic content. This is an important finding for Technology intelligence, since a larger number of internally coherent topics is an extremely useful starting point for the interpretation of their content.

Second, in the case of the Functional dictionary the Document-Feature Matrix (DFMs) is less sparse: the average sparsity index is 73,8%, as opposed to around 98% for the non-technical dictionaries and 94,3% for the Advantages & Disadvantages dictionary. Third, the total number of documents used, i.e. the documents in which there is at least one matching between the words and the features of the dictionary, is larger in the case of technical dictionaries (in particular, it is 498 out of 500 for the Functional dictionary).

Fourth, the cosine distance of topics is approximately ten times larger in the case of technical dictionaries, meaning that clusters enjoy a ower cohesiveness. These findings point to an important contribution of technical dictionaries, and in particular of the Functional dictionary, to Technology intelligence, that is, technology mapping, clustering, interpretation, and foresight.

The Functional dictionary offers a large number of features that find a matching with the text of the patents, so that the resulting DFM is less sparse. The fact that the total number of topics is larger also means that each of the topics is populated by a smaller number of features. Each topic, therefore, lends itself to a clear understanding of the technical content. The differences between topics have also a clear technical interpretation.

It can be said that technical dictionaries, as opposed to non-technical dictionaries, allow a fine-grained intelligence of the technical content of patent sets. This is in itself an important achievement, given that the topic modelling algorithm has not been previously trained, but is only filtered by the dictionary. In practice we achieve the precise results of supervised topic modeling with an effort which is more similar to that of the unsupervised approach.

In order to give a better understanding of the potential of the technical dictionary approach, let us examine more in detail in table 12.4 and 12.4 the top 10 words in some of the topics identified in one of the four patent sets examined, namely Prosthesis.

Table: Top ten words in selected topics extracted in the patent set **Prosthesis** with the aid of the **Advantages & Disadvantages dictionary**.

Topic 6: Term	Topic 6: Beta	Topic 28: Term	Topic 28: Beta
Adjustable	0.564	Worn	0.155
Accommodate	0.046	Comfort	0.086
Ergonomic	0.032	Wearing	0.077
Securely	0.022	Comfortable	0.058
Adjustability	0.020	Injury	0.043
Ability	0.017	Discomfort	0.031
Facilitate	0.016	Increase	0.029
Stabilizing	0.011	Reduce	0.026
Comfort	0.010	Uncomfortable	0.025
Stability	0.010	Tear	0.024

Table: Top ten words in selected topics extracted in the patent set **Prosthesis** with the aid of the **Functional dictionary**.

Topic 15: Term	Topic 15: Beta	Topic 20: Term	Topic 20 :Beta
Distance	0.054	Arm	0.238
Coupling	0.053	Exercise	0.082
Position	0.052	Arms	0.060
Neck	0.043	Stretching	0.042
Line	0.041	Left	0.025
Parallel	0.030	View	0.024
View	0.021	Orientation	0.023
Form	0.019	Stretch	0.021
Orientation	0.016	Angle	0.019
Contact	0.013	Tension	0.017

In Table 12.4 Topic 6 calls the attention to features of the prosthesis such as adjustability, facilitation and stability, that is, on ergonomic features. Topic 28 emphasizes comfort (as an advantage made possible by the invention) or discomfort (as a disadvantage addressed by the invention). The content of these topics, thanks to the power of the technical dictionary, is crystal clear.

In Table 12.4 we see, in topic 15, the positioning of prosthesis on the neck of patients, and in topic 20, the orientation of arms with the support of prosthesis. Both these topics deal with the issue of placing the prosthesis around or over various parts of the human body. It is important to underline that, once generated, topics can also be combined together in further steps of the analysis.

Blacklist

Finally, the information power of technical dictionaries can be appreciated by comparing the results with those that would be obtained by filtering the texts with semantically poor words, such as generic words (stopwords) and "patentese" generic words. As stated above, the combination between generic and patentese stopwords is labelled Blacklist.

Table: Top ten words in selected topics extracted in the four patent sets with the aid of the **Blacklist**.

Dataset	Number of topics	Cosine Distance of Topics	Number of documents	Number of features	Average sp
Burners	40	0.016	500	500	
Toothbrushes	40	0.025	500	500	
Prosthesis	40	0.040	500	500	
Funeral Dev.	40	0.042	500	500	
Mean	40	0.031	500	500	

The Table 12.4 shows the clustering results using as filter a blacklist, composed by traditional and patentese stopwords. The following observations are in order. First, we observe a remarkable flatness in the number of topics: it equals 40, which is the maximum value that can be achieved by the algorithm. Second, Burners and Toothbrushes have a smaller distance value, confirming that these two patent sets are composed by more homogeneous types of features.

Summing up, it seems that the dictionaries deliver different results. On the one hand, Non-technical dictionaries include more generic word expressions. Non-technical dictionaries deliver better performance if the goal is to provide a global representation of the technological field, since they generate a smaller number of topics and better indicators of effectiveness of clustering.

On the other hand, if the goal is to detect novelty and technological trends, or to identify the areas of exploration of emerging technologies, then Technical dictionaries perform better, as they generate a larger number of topics. Analysing these topics it is possible to get insights on technological differences among patents in the same patent set and to produce detailed technology maps. This is particularly important for innovations in the pre-paradigmatic stage.

12.5 Conclusions

Patent clustering is a standard application in technology intelligence and has been common practice in the professional IP industry. In recent years, patent clustering based on text mining has gained large acceptance. In this paper we have introduced a novel methodology to cluster patents. With respect to the existing literature we give a contribution by testing the use of dictionaries as a structured list of words to filter patent data and build up clusters. The clustering based on dictionaries is significantly different from the one provided by IPC classification. We have tested the use of multiple dictionaries on the same patent set. The resulting clusters allow a fine-grained interpretation, which is illuminating for purposes of technology intelligence. Kreuchauff and Korzinov (Kreuchauff and Korzinov, 2017) have developed a set of performance criteria to compare and evaluate the identification approaches proposed in the literature. These criteria include: degree of intervention of experts, portability, transparency, replicability, adaptability, updating capacity, and finally extent and relevance of data obtained. We suggest that the enriched dictionary approach satisfies all these criteria: it does not involve the role of experts, is portable and transparent (after publication of the dictionary in the open literature), it is therefore replicable to any technology, is adaptable and has capacity to update (particularly if based on Wikipedia), and offers broad and relevant data.

Chapter 13

Impact of Research from the Perspective of Users.

A substantive interest has been developed in the last 15 years on the so-called "impact revolution", namely the increasing demand for showcasing the results of publicly funded research in order to justify public expenditure. Public funders are increasingly required to demonstrate the relevance of funded research not only for scientific communities but also for the economy and society at large. In other words, there is an increasing demand to prove that the users and beneficiaries of research results are not only the traditional academic audience - researchers and university students - but include a large number of social actors. Let use use here the concept of "societal impact", as opposed to academic impact, to include all dimensions of impact on the society and economy that are realized through impact pathways that go beyond the institutional research and teaching activities.

The issue of societal impact of public research has gained prominence in the specialized literature since the start of the century and has accelerated in recent years (Van der Meulen and Rip, 2000; Ernø-Kjølhede and Hansson, 2011; Bornmann, 2013; Bornmann and Marx, 2014; Bornmann and Haunschild, 2017). Indeed, a quick look at the most important journals in the field of innovation, research policy and research evaluation shows that the most largely cited, downloaded or read articles in the last five years are almost invariably dedicated to the issue of societal impact.

This follows from the adoption of societal impact of research as one dimension of evalution of research, both ex ante and ex post, in many advanced countries.

As discussed by several authors, societal impact has become one of the criteria of ex ante project selection in several institutions and countries (Kanninen and Lemola, 2006; Dance, 2013). It is also a crucial chapter in the ex post research assessment in some countries, such as United Kingdom. Within the UK Research Excellence Framework (REF) the assessment of societal impact has been responsible for 20% of the total score, while an increase to 25% has been announced in September 2017 for the future exercise. The publication of REF case studies of societal impact has fueled a field of analysis (Derrick et al., 2014; Samuel and Derrick, 2015; Khazragui and Hudson, 2014). Some authors advocate impact analysis as a way to examine the effects of research agenda on the societal priorities (Cozzens et al., 2002).

This surge of policy interest, however, comes in a period in which the scientific analysis of the concept of societal impact and of the potential and limits of existing methodologies has not yet come to a general agreement. As succinctly stated by Lutz Bornmann, impact evaluation is "still in the infant stage" (Bornmann, 2013). And Bozeman and Sarewitz (Bozeman and Sarewitz, 2011) explained that "there has been remarkably little progress in the ability to measure directly, systematically, and validly the impacts of research on social change", so that "we have no satisfactory analytical tools for characterizing the social impact (of research)" (Bozeman and Sarewitz, 2011).

This paper is a contribution to the substantive and methodological work on the assessment of societal impact

of research. From the substantive point of view, it introduces the notion of target group, or group of potential users of research, as a necessary component of the design and implementation of research projects. From the methodological point of view, the paper strongly supports the idea, already advanced in the literature, that Text mining techniques are promising in this field, but suggests a major modification by introducing the Enriched dictionary methodology.

To be more precise we argue that a necessary component for impact assessment is the definition of users of research at a granular level. In addition, we suggest that the more researchers are able to define precisely their target groups the more they are likely to reach them effectively and to increase the impact. We develop a full scale, replicable and scalable methodology to identify the user groups mentioned in research-based texts, such as research proposals (ex ante), impact case studies (ex post), or publications. We test the methodology on the collection of case studies developed under the Research Excellence Framework (REF) in the United Kingdom. We examine three main dimensions of user target groups (frequency, intensity and specificity) and disaggregate the data by broad discipline.

13.1 Methodological challenges

13.1.1 Variability in the identification of outcomes and users

A first methodological issue is that in order to assess the societal impact of research, there is a need not only to identify observable elements that can be considered as an outcome of the research process, but also to define the actors that are affected, or benefit, from those outcomes. It turns that these elements are subject to huge variability across disciplines. Consequently, there are areas in which methodologies are more sophisticated and largely tested, and others in which there is remarkable lack of experience and methodological work (Stern et al., 2013; Mitton et al., 2007, Ţurcan (2015))

Among the former the health care sector is probably the one in which the impact assessment of research has made the largest progresses: a number of well structured research impact assessment methodologies have been developed and implemented. There in fact are as many as 16 different impact assessment models, according to Milat, Bauman and Redman (Milat et al., 2015). An important reason for this accumulation of experience and knowledge is that the outcomes that demonstrate the expected impact are clearly identified and standardized, and the categories of users are clearly observed, given a high degree of professionalism.

At the opposite side of the spectrum, there is still much uncertainty about the way in which the societal impact of research in social sciences and humanities (SSH) could be defined and observed, even less quantified and measured. The challenges associated with identifying the impact of outputs from these fields stem from a number of issues, most of which have been noted in earlier evaluation-based literatures. According to certain theoretical perspectives the very notion of impact is problematic for SSH (Blasi et al., 2018). From a historical perspective, it has become increasingly clear that research across SSH has had a large influence on modern societies on a long time scale (Bod, 2013). It should be recognised therefore that a certain share of research need not be asked to demonstrate any impact, but be valued for its own sake (Small, 2013). It is part of the millennial history of humankind that some people, some ages of life, some resources are dedicated to the search for intangible and priceless goals such as beauty or truth. Research from the arts and humanities is needed in order to preserve in society the ability to interpret, appreciate, enjoy and valorize symbolic values inherited from the past. Should the many scholars from this field be interrupted or deprived, modern societies would rapidly become unable to coordinate, administer and govern themselves.

Consider the problem from the perspective of potential users of SSH research: the results or products are not necessarily used on the basis of direct access to scientific sources (as it happens more frequently with technological and biomedical research), but after some transformation and intermediation by specialised actors (e.g. journalists of popular magazines; social media). Furthermore they do not necessarily take the form of compelling evidence, or ultimate scientific authority, but enter into a social arena for public and political conversations and debate, where arguments may be advanced and refuted. In addition, audiences may be dispersed, non-institutionalized, or even transient (e.g. issue-based) and not professional. Finally, social behaviors are by definition slow to change, thus the impact of research is likely to be seen only after a

long delay. This means that both outcomes and categories of users are much more difficult to identify and define.

13.1.2 Sources of information

Current approaches can be classified, according to Morton (Morton, 2015), as forward tracking, backward tracking, and evaluation of mechanisms. In forward tracking, researchers are asked to reconstruct the ways in which their research might be useful for given categories of users. Alternatively, users are asked to declare which kind of research results they are likely to utilize (Tang et al., 2000). The strong limitation of this approach is that it relies heavily on the researcher's and research user's own recollections of research use (Nutley et al., 2007; Donovan, 2011). In some sense, this is also the limitation of the use of case studies of research impact: it is difficult to verify whether they are a random collection or they are biased in one way or another. Backward tracking suffers less from these subjective biases. If the sample of final outcomes is well designed, it can offer important lessons for researchers and policy makers. However, it comes with long delays with respect to actual research results.

It should be noted that this methodology is the one largely adopted in impact assessment of health-related research: once it is agreed that clinical guidelines are a suitable candidate for assessment, it is possible to trace back the impact using the citations to the medical literature. Evaluation of mechanisms is a partial methodology, which describes in great detail the pathways in which research results are channeled from their origin to the endpoint.

13.1.3 Text-based impact assessment

More recently, an interesting alternative approach has been suggested. Based on the availbility of Machine learning and Text mining techniques, it has been argued that the evidence for the impact of research might be traced by extracting selected expressions from certain kinds of documents. We may distinguish between two kinds of documents: (a) produced by users of research; (b) produced by researchers themselves. There are several suggestions to use documents produced by users of research. In one of the most developed efforts to conceptualize the social impact of research, called Public Value Mapping (PVM) Bozeman and Sarewitz (Bozeman and Sarewitz, 2011) suggested the use of three main sources of statements, from which it could be possible to trace the impact of research: government statements, academic literature, and public opinion polls containing public statements. More recently, Bornmann, Haunschild and Marx (Bornmann et al., 2016) have suggested to use the frequency of occurrence of policy-related words in policy documents as evidence of impact of research. In this paper we will make these suggestions operational.

Yet another approach in the same line is to examine the documents produced within social media. A prominent approach is based on Altmetrics measures. The main tenet of Altmetrics is that citations, the basic unit of analysis of bibliometrics, capture only part of the impact of published research, so that "citation tracking has never been able to follow the less visible- but often more important- threads of invisible colleges, woven through personal connections and informal communications" (Priem et al., 2012). By accessing data on the personal use of published materials "Altmetrics could deliver information about impacts on diverse audiences, like clinicians, practitioners, and the general public, as well as help to track the use of diverse research products like datasets, software, and blog posts" (ib.). Sibele et al. (Fausto et al., 2012) examine in this light the phenomenon of research blogging. There are however severe limitations that might make Altmetrics problematic. The extent to which Altmetrics can capture traces of societal impact has recently been seriously contested (Bornmann and Marx, 2014). Social media are used more for internal discussions within scientific communities, rather than a bridge between the research community and society at large. According to Haustein there is lack of evidence that social media events can serve as appropriate indicators of societal impact (Haustein et al., 2016).

Among the documents produced by researchers, we might further distinguish between: (a) research proposals (ex ante documents) and (b) case studies produced after the realization of research projects (ex post, or documents produced within the research assessment process). It is this type of documents we suggest to

examine as a novel methodology to assess the potential societal impact of research. In this paper we follow type (b) documents and use the collection of case studies produced by UK researchers under the REF exercise. In future studies we plan to use archives of proposals made available in public sources in order to examine the ex ante representation of researchers. The REF impact case studies, as already noted, have been the object of a large literature in recent years. Among these studies, King's College and Digital Science (London and Science, 2015) have indeed produced, using Text mining techniques, an interesting analysis of beneficiaries of UK research, publishing a fascinating infographics. This analysis, however, lists only a fairly small set of research users, most of which are defined with generic terms. It is our contention that much further work should be done in this direction. We propose a new lexicon-based methodology, called Enriched dictionary (see below), which allows a much more fine-grained analysis.

13.2 Methodology

Basically we suggest to examine carefully the full text of documents produced by researchers and extract, in a highly structured and theory-dependent way, information on potential users of research. Users are defined as categories of human agents that share some characteristics that are relevant with respect to the object of interest. In the present context users are social groups, or target groups, that are potentially affected by research results and that use these results for their own purposes. Before entering into a technical description of the methodology let us address the rationale of assuming users as an important dimension of research impact.

There are several compelling theoretical reasons for this choice. First, the literature has strongly underlined the interactive nature of societal research impact. As discussed above, the most recent literature and practice strongly suggest to abandon a unilinear model of impact, in which it is expected that researchers produce results, diffuse them in various channels, and see the results taken up by interested users. Let us call this approach a "percolation model": researchers produce results that eventually percolate down into society, but without knowing the ways in which they flow, the obstacles they meet, the timing of the process, or the final destination of the flows. On the contrary, it is strongly suggested to adopt an interactive model of interaction, in which researchers actively engage into systematic relation with potential users. In an interactive model there must be a reflexive activity on one side about the nature (characteristics, interests, behavior, style) of the other side. Researchers must build up a representation of their potential users, and vice versa. How could researchers engage with users if they do not know them? And how they could know potential users if they do not engage into some sort of analysis, even as simple as description and characterization? For interaction to take place, there must be some preliminary recognition of the existence, nature, attitudes of those that may utilize the research results. According to our methodological suggestion, it is this representation that is the preliminary object of interest for impact assessment. If researchers have a representation of their potential users, they will leave traces of this representation in their written texts. When they write research proposals they will promise to address the issues of these users, and when they write case studies of research impact they will report on the takeup or use of their research activities by these users. Second, it has been shown that research activities have a huge variety of impact pahways, largely dependent on the scientific discipline. In turn, this implies that disciplines have at least partially different target groups. Research in political science is different from research in oncology not only because their scientific foundations, methods, objects and cognitive styles are different, but also because they talk to different user groups. The texts produced by researchers themselves are a necessary starting point to reconstruct the various impact pathways. Third, focusing on user groups has the advantage of shifting away the attention from discrete events or products to long term processes of interaction between research and society. The focus on discrete events or products is a typical feature of the narrow definition of research impact cultivated since long time in the so called "valorization of research". This impact is defined and measured with reference to highly stylized entiities, such as patents, licensing contracts, research contracts, and spinoff companies. These are clearly defined, legally enforced, highly visible and measurable entities. Defining and measuring impact is easier by focusing on these entities because they convey the meaning of knowledge transfer from research to the market, and because the final outcome can be defined in monetary terms. We suggest to focus on user groups as a relatively permanent social entity, which is defined by a specific combination of social status, needs, culture,

practices and routines. User groups survive the individual personality of people. They are a permanent, although often entirely informal, characterization of society. Finally, our methodology allows the large scale automatic analysis of large corpora. This means that the inevitable subjectivity in the reconstruction of impact by researchers in writing their proposals and/or impact case studies can be mitigated by examining large scale patterns. It is important to remark that the notion of users is consistent with other suggestions in the literature that adopt different definitions, such as stakeholders, constituencies, interest groups. Our definition is broader and admits more internal variability, as discussed below.

13.2.1 Operationalizing user groups using Natural Language Processing techniques

A simple implication of our methodology is that researchers "leave a trace" of their representation of users, or the groups of social agents that are most likely to use or uptake the results of their research activity. By using state-of-the-art Text mining technologies we are able to identify these traces in written texts and to give them unambiguous meaning. By assuming target groups as units of analysis we suggest to introduce a number of concepts, from which suitable indicators can be derived.

Definition 1 Stakeholders are entities influenced by the research activity. This definition covers all possible entities that engage an active or passive relation with the research activity.

Definition 2 Target groups are entities or groups of entities on which researchers claim to have an effect.

Given definition 2, it is clear that every target group is also a stakeholder, while the reverse does not hold true. Non-target stakeholders include the proponents themselves, managing autorities, funding agencies and so on. We need a formal technique for identifying target groups in the text of research documents. This technique as been developed as described in section 7.1.

Dictionaries are a peculiar type of written text, characterized by authoritativeness, saturation and update. In other words, a dictionary must be composed of entries established by some authority, most often an academic one and/or an authority established since long time by reputation (e.g. editorial initiatives of prestigious publishers). Saturation means that all words that are related to the domain of the dictionary must be included. It is a major flaw of a dictionary the lack of important entries. A dictionary is characterized by a property of semantic saturation: all words that have a meaning associated to a given field are included in the dictionary. Using this tool it will be possible to count the occurrences of target groups, and develop indicators of frequency and intensity. Finally, update means that dictionaries have an internal organization (for example, an editorial board) that examines all new expressions, discusses their acceptability in the dictionary, and make official and authoritative decisions about inclusion or exclusions.

These formal requisites, that used to be appropriate only for established dictionaries, are currently satisfied by a larger variety of sources. In particular, the huge power of Text mining techniques has made it possible to automatize at least some of the steps needed to create a formal dictionary. Section 7.1 illustrates the steps undertaken in order to build up an Enriched dictionary of users. It currently includes 76.857 entries, that have been shown to saturate the semantic field of users. It includes, among others, all jobs, work positions, professions, hobbys, patient roles, sports, creative and enterteinment roles, political, institutional and organizational roles, social roles, that have been classified in hundreds of official sources. In particular, this includes all stakeholders and target groups, as defined above.

Our Natural Language Processing (NLP) system follows the following steps. - Sentence splitting and Tokenization: this process splits the text into sentences and then segments each sentence in orthographic units called tokens. Sentence splitting plays a key role since thanks to a given word, it is possible to find all sentences in which the word is used. - POS tagging and Lemmatization: The Part-Of-Speech tagging (or POS tagging) is the process of assigning unambiguous grammatical categories to words in a specific context. It plays a key role in NLP and in many language technology systems. Once the computation of the POS-tagged text is completed, the text is lemmatized according to the result of this analysis. - Target groups Annotation: The Target groups Extraction tool is based on lexicon methods. Among the various lexicon methods we adopt, as stated above, the Enriched dictionary approach. With respect to users, we use a lexicon composed

of 76.857 entries. By launching this Extracion tool we are able to capture all the different ways in which each target group can be expressed in a research document.

Table 13.2.1 shows the output of the NLP procedure for a sentence contained in the corpus ("Each year, in England alone, approximately 152,000 people suffer a stroke."). As it can be seen, the automatic annotation system isolates the only word ("people") that may be part of a target group.

Table: Tokenization, lemmatization and annotation of a sentence in the corpus.

Doc_id	Sentence_id	Token_id	Token	Lemma	Xpos	Full_target_group
1855	1	1	Each	Each	DT	NA
1855	1	2	year	Year	NN	NA
1855	1	3	,	,	,	NA
1855	1	4	in	in	IN	NA
1855	1	5	England	england	NN	NA
1855	1	6	alone	alone	RB	NA
1855	1	7	,	,	,	NA
1855	1	8	approximately	approximately	RB	NA
1855	1	9	152	152	CD	NA
1855	1	10	people	people	NN	People
1855	1	11	suffer	suffer	VBP	NA
1855	1	12	a	A	DT	NA
1855	1	13	stroke	stroke	NN	NA
1855	1	14	.	.	.	NA

13.3 From text extraction to indicators

After having extracted all possible expression of target groups in research documents, we are in a position to develop indicators with suitable statistical properties. They are defined as follows.

Frequency

For a given document J, let us define T_j (number of target groups contained in J) and W_j (number of words contained in J). We then define:

$$F_j = \frac{T_j}{W_j} * 100$$

The frequency F of a document measures the percentage of words that are target groups. If a document shows high frequency it means that it cites many times target groups, even if it/they are always the same and they are generic. For example, an impact description that repeats many times the target group people will show high frequency.

Diversity

For a given document J, let us define Tu_j (number of different target groups contained in J) and Wu_j (number of different words contained in J). We then define:

$$D_j = \frac{Tu_j}{Wu_j} * 100$$

The diversity D of a document measures the percentage ratio of different words that are target groups. If a document has a high diversity it means that it cites many different target groups, even if they are generic.

Specificity

For a given target group i, let us define N (number of document in the corpus) and ni (number of documents that contains the target group i). We then define Si, the specificity of the target group i as:

$$S_i = \frac{log(N/n_i)}{log(N)}$$

The Specificity of a target group Si measures how rare, and thus specific, is a target group in the overall corpus. The specificity diminishes for target groups that occur very frequently in the corpus and increases for target groups that occur rarely (more specific target group). Let us take the example before, in which the annotation system identifies people as a target group. There will be a large amount of documents in which the word people will occur, so that the ratio between the total number of documents in the corpus and the number of those that include the word people will be close to one. On the contrary, highly specific words (say, free climbers) will occur less frequently, so that the above ratio will increase. Since we are interested in giving a measure for each document, having defined Si for each target group, if the document j contains k different target groups, we have that the specificity of the document j is:

$$Sj = \sum_{i=1}^{k} Si, j/k$$

The specificity of a document Sj measures how rare, and thus specific, are the target groups contained in that document and it is the mean of the specificity of all the target groups that it contains. If a document contains only rare target group (not cited by other impact descriptions) that document exhibits high specificity. For the previous example, suppose we have a document repeating many times that the research has an impact on people. Since the target group people is a common one (thus having a low specificity Si itself) the measure of the specificity of the document, resulting from the sum of low specificity values, will be low.

An interesting example of application of these principles comes from the scientific study of popular science, or divulgation. In these fields authors use a language which must be understood by lay people, not by professional scientists. For this reason they tend to use generic words, rather than highly specific and professional words. This is, incidentally, one of the reasons why professional scientists ofte disregard popular science as a literary genre: they perceive the generic nature of language used as too coarse. Scholars of popular science have used quantitative linguistic techniques to distinguish between generic and specific terms in the same semantic field (Jacobi, 1999)

13.3.1 The meaning of Frequency, Diversity and Specificity indicators for the analysis of research impact

The above definitions are building blocks of a model of engagement of researchers with their potential users. At the outset, it is important to examine whether researchers include users in their representation of research activity at all. Thanks to the use of an Enriched dictionary, which by definition saturates the semantic field, we are in a position to establish whether user-related expressions are found in researchers' texts or not. Since in this paper we use the impact case studies produced under the REF, the minimum level is by definition satisfied, as it was the condition for submitting the case study. At the same time, this indicator will be extremely useful for the ex ante evaluation of research proposals. The appropriateness of this indicator and its policy implications will be the object of future research. Authors of the REF documents are by definition aware of the existence of target groups of users.

Once the awareness level is satisfied, frequency comes into play. Frequency is a standard measure in computational linguistics and Text mining techniques, since it gives evidence of the relative importance of words or expressions. The frequency by which target groups are mentioned in a document is a measure of their perceived importance. We are keen to examine the frequency by which users, or target groups, are mentioned in documents produced by researchers. Yet researchers may cite repeatedly a target group, but consider them

as a unique entity. This approach is reasonable when the target group does not have internal differentiation (i.e. it is not segmented) and when the results of the research are equally useful for all its members.

But in many relevant cases this undifferentiated approach does not work. Representing and addressing users as a single target group may weaken the potential for impact.

There are two directions in which researchers can deepen their representation of target groups, and hence their impact. One is to address different target groups. This is similar to the notion of segmentation in consumer psychology: if target groups have sufficiently dissimilar characteristics with respect to the activity (in this case, the use of reserch results), then it is better to treat them differently. The other is to go in depth in addressing each of the target group, by refining their approach, using fine grained representations of the needs of the users. We capture these two directions by measuring diversity and specificity. It is our contention that the language used by reseachers is a clear signal of their approach to research impact. A high level of diversity implies that researchers understand the need to mention, identify, enumerate target groups that have different names. They stop using generic words (say, people) and start to introduce some of the many criteria for segmentation. In addition, or in alternative, they discover that within their target groups it is possible to go deep in the fine grained representation, by adding specificity.

A spatial metaphor may help to capture the point: by increasing diversity of target groups researchers move horizontally, defining new regions of the space, while by increasing specificity they move vertically, drilling the ground in each of the regions. Frequency, diversity and specificity are not necessarily correlated. An interesting empirical issue is the relation between these two dimensions.

Let us articulate an example. Suppose an expected impact of a given research is on policy making. A low specificity situation arises when researchers speak about "policy makers", or "government". A more mature and engaged approach should articulate the policy making process by identifying several specific user groups in addition to the various layers of political and legislative decision making. For example, interest groups. These social actors are extremely important in shaping the policy agenda. A well developed body of research in political theory has examined the way in which new policy issues are generated, framed in the public conversation, and pushed forward in the policy arena until they become established in the policy agenda (Sabatier, 1987). Pittman (2006) argues that the most important factors leading to government interest there is the role of domestic advocacy, as well as the interest in their international standing. Second, intermediary organizations or boundary organizations, such as technical agencies and regulatory bodies (Agrawale, Broad and Guston, 2001). Third, opinion-makers such as think tanks should be included. Many other examples could be added. These actors would be mentioned with more specific expressions.

Finally, researchers may engage into identifying user groups "by name and surname", that is, as concrete and localized actors with whom they plan to enter into interaction. Counting or measuring them would be the final stage of maturity of research engagement with users.

13.4 Data

13.4.1 Description of the corpus

The corpus is composed of 6637 REF impact case studies. They generally follow a template illustrated in the REF criteria. The template has a Title and five main text sections, plus the name of the Submitting Institution and the Unit of Assessment. In addition to the Title of the case study, the text sections of the template and the indicative lengths, as recommended in the REF criteria are:

1. Summary of the impact, 100 words
2. Underpinning research, 500 words
3. References to the research, 6 references
4. Details of the impact, 750 words
5. Sources to corroborate the impact, 10 references

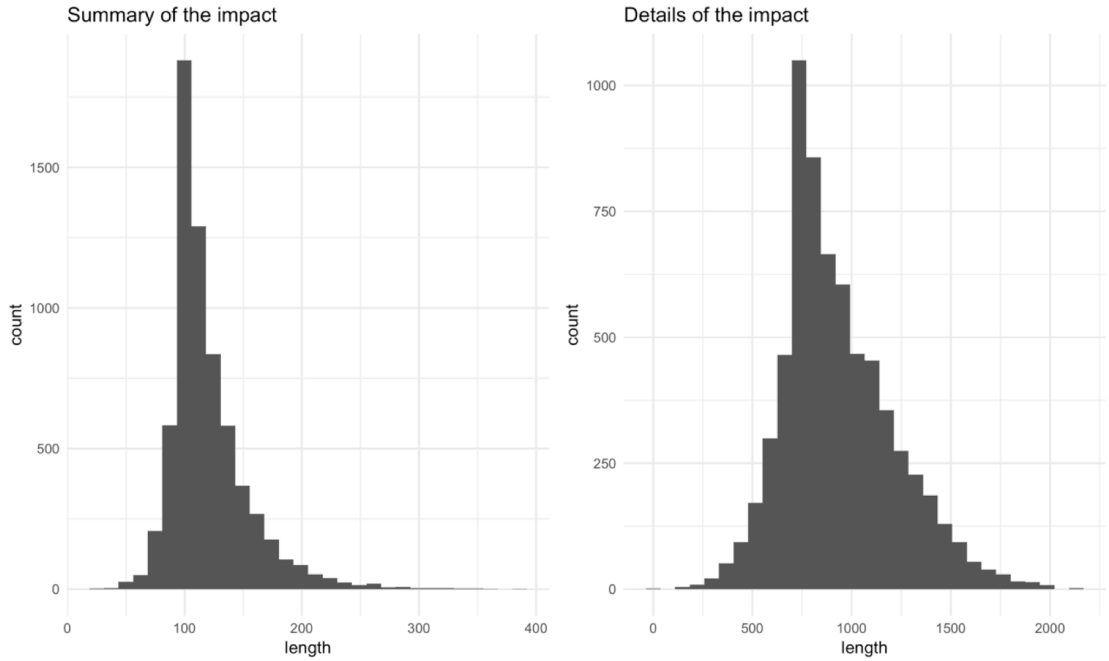

Figure 13.1: Distribution of number of words in relevant sections of the REF impact case studies.

We take into consideration the sections Summary of the impact and Details of the impact. It is common practice in computational linguistics to examine the length of documents to be included in a corpus in order to ensure comparability. Figure 13.1 shows that the limits established by the REF criteria are not always respected. Nevertheless, since the distribution of the length is almost normal and there are not outliers it is appropriate to include all documents in the corpus.

Within the REF repository projects are classified using three criteria:

- Impact type: There are eight Summary Impact Types. These follow the PESTLE convention (Political, Economic, Societal, Technological, Legal, and Environmental) widely used in government policy development, with the addition of Health and Cultural impact types.
- Units of assessment (UOA): Institutions were invited to make REF submissions in 36 subject areas, called units of assessment (UOAs), each of which had a separate expert panel.
- Research subject areas: The REF Impact case studies are assigned to one or more Research Subject Areas (to a maximum of three) by text analysis of the 'Underpinning research' (Section 2 of the Impact case study template). This is a guide to text search that uses a disciplinary structure that is more fine-grained than the one in the 36 Units of assessment.

Figure 13.2 shows the number of documents per Unit of assessment.

13.4.2 Preliminary analysis of the corpus

In this section we present a descriptive analysis of the content of the documents to give an evidence of two important facts: 1. The description of the impact contains target groups; 2. This information is enough to make a significant statistical analysis.

The corpus contains 8.230.598 words in total and 141.705 different words. By annotating the entire corpus with the entries of the Enriched dictionary we find that the total number of words referring to target groups is 169.037, while the number of different target groups is 1830, or 1.3% of different words.The number of

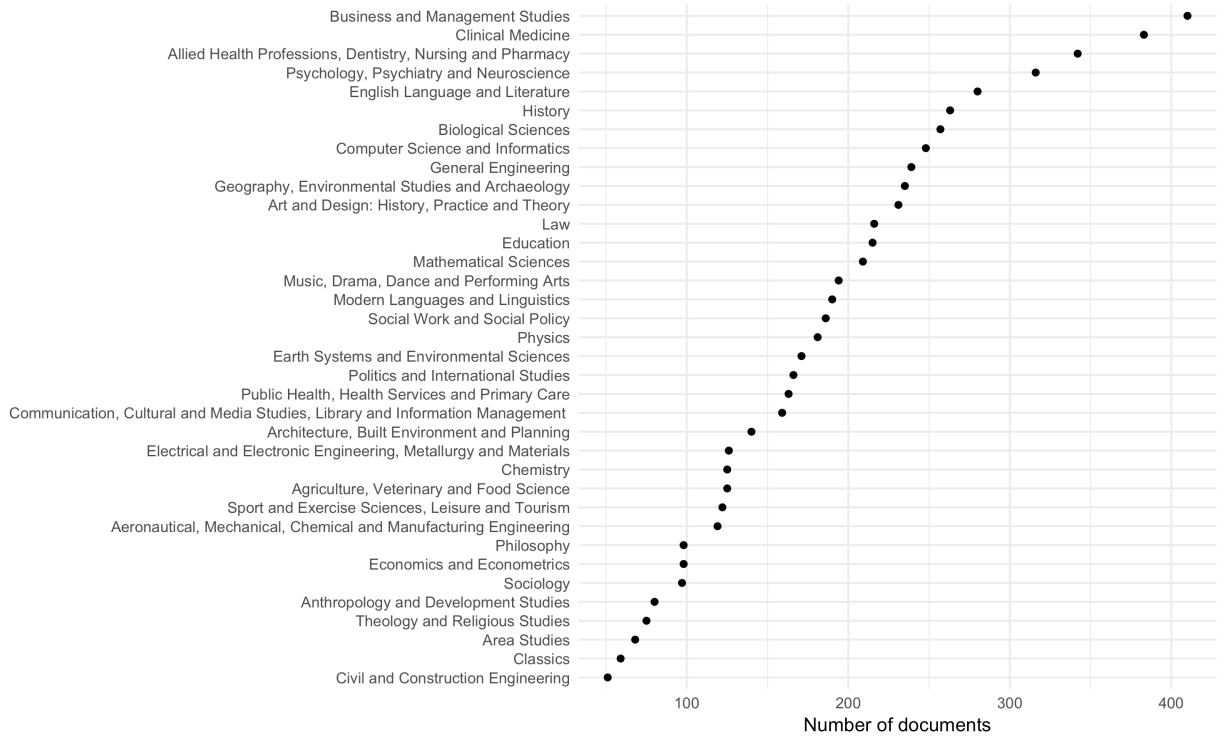

Figure 13.2: Number of documents per Unit of assessment (UoA) in REF impact case studies.

documents that contain at least one target group is 6628, or 99,9% of the total. Only for nine documents we were unable to locate any word referring to a target group.

Figure 13.3 offers a vivid demonstration of the issue of specificity of words referring to target groups. As many as 37% of all projects include people, and as many as 25% mention company as isolated words. Among the top 20 occurrences we find extremely generic words such as public, community, individual, organization, user, or society. Slightly more specific are the words referring to the school or youth context (child, school, student, teacher) or the health context (patient, patients). In order to find more specific words we have to go further down the ranking. Please note that in all these cases these words do not appear in combination with other that might increase the specificity, but in isolation. Should the same word appear in combination with other more semantically connotated words, they would form a separate target group. As an example, the word people is considered part of a separate expression in the following examples: people with cystic fibrosis, people with primordial dwarfism, people with rheumatoid arthritis, ordinary people in extraordinary situation, people in senior management, people from different background, key policy people in uk government, specific community of people, young people in deprived community in Glasgow. Each of these expressions is considered as a separate target group and their Specificity is computed according to the formula above.

To see the difference in the use of words in sentences consider the following examples from REF case studies: "The validation of the exit poll forecast allowed people to see the power of social scientific methods, and may have helped them to establish a level of trust in evidence-based information" (generic use of the word people) and "Key components of this are nurturing people with cross-cultural understanding, diversity in thinking and global leadership skills" (more specific use). The NLP procedure developed for this analysis is able to accurately distinguish these situations.

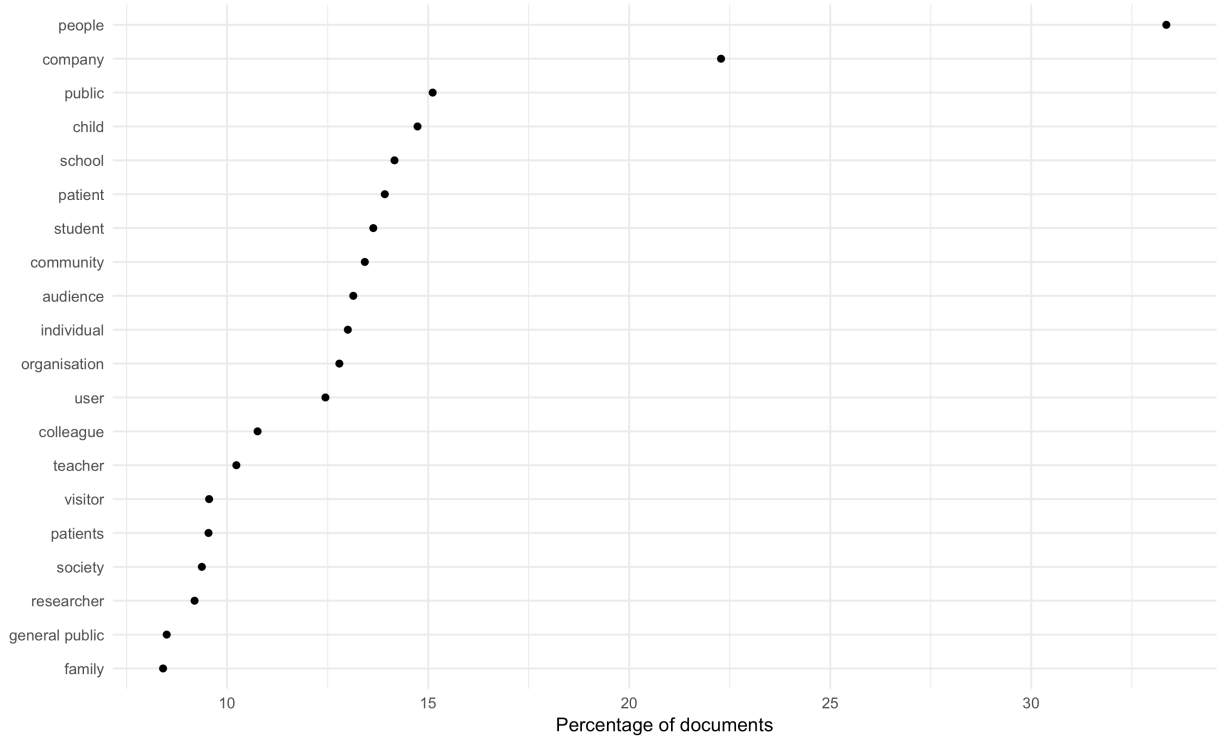

Figure 13.3: Top 20 occorrences of words referring to target groups in the corpus of REF impact case studies.

13.5 Results

13.5.1 Descriptive analysis

Table 13.5.1 offers a snapshot of the value of indicators calculated across the entire corpus. The minimum value of indicators (zero) is represented by the 9 documents without target groups. On average the REF impact case studies contain words that represent target groups as 2% of total words, and words that represent distinct target groups as 2.5% of different words. In absolute terms, the median REF impact case study contains 10 different words thet refer to target groups, repeated 22 times in total.

Table: Descriptive statistics of indicators of target groups in the corpus of REF impact case studies.

Indicator	Minimum	Maximum	1st quartile	3rd quartile	Mean	Median
Frequency:indicator	0	9.2	1.2	2.7	2	1.9
Frequency: absolute value	0	115	14	34	25469	22
Diversity: indicator	0	7.6	1.8	3.2	2.5	2.4
Diversity: absolute value	0	34	7	14	10518	10
Specificity	0	1	0.59	0.731	0.659	0.656

There is not large variability in the number of different target groups, as the first and third quartile at 7 and 14 are close to the median value. Interestingly, all distributions are close to the normal. A small number of documents use words that refer to target groups with much larger frequency and diversity. The document with maximum use of target groups identifies as many as 34 different words or combination of words. In terms of repetition, the document with the largest number of words uses 115 times a word representing target groups. When coming to specificity, the mean value of 0.659 implies a good level of specificity, given the range of the indicator. Figure 13.4 shows the distribution of documents by indicators.

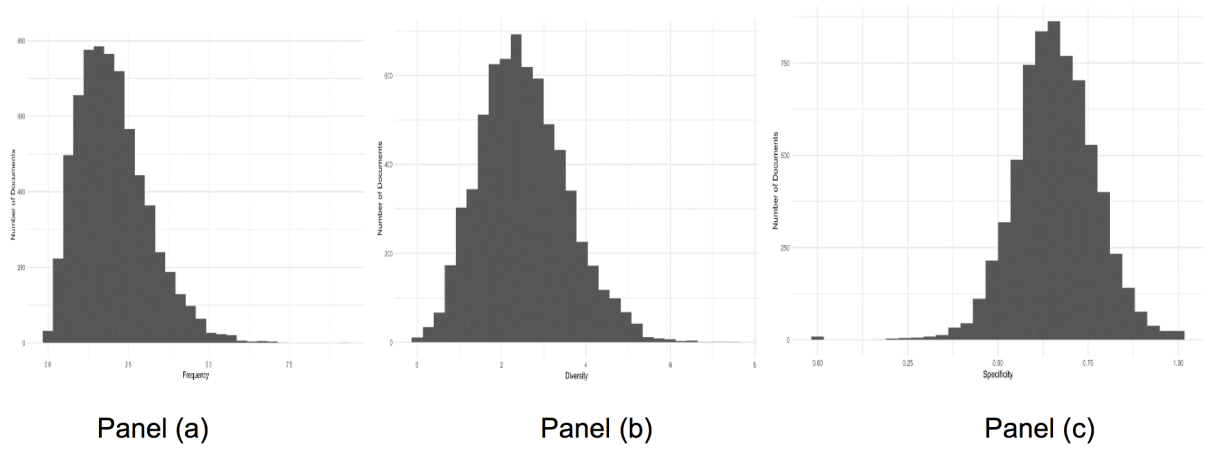

Figure 13.4: Top 20 occorrences of words referring to target groups in the corpus of REF impact case studies.

13.5.2 Findings by subject area

We disaggregate the indicators with respect to the subject areas, or the Units of assessment (UoA), according to the REF nomenclature. This analysis offers a new perspective on the way in which he various disciplines describe their impact pathways. We use the conventional boxplot representation.

Figure 13.5 shows a surprising finding. On top of the ranking by frequency of occurrence of words that refer to target groups of users we find Humanities: not only Education (which by nature refers to children, students and teachers as user groups), but also Music and drama, Classics, Language and literatire, and History. Or, in other words, disciplines that are not oriented towards users, but rather cultivate the goal of knowledge per se. At the bottom of the ranking we find, again surprisingly, Engineering disciplines, with a number of specializations, and Economics, that is, disciplines that, on the contrary, have a pragmatic orientation towards various target groups of users.

Figure 13.6 confirms a similar ranking by subject area when we use the Diversity indicator, with slight variations. In the case of diversity we find also Theology and religious studies, as well as Art and design. Almost all subject areas in Humanities consistently show up at the top when we investigate the frequency by which they mention users of research and the number of different target groups they are able to identify. This is a remarkable finding. It is true that this comes from documents that are themselves retrospective reconstruction of impact, but this feature applies to all subject areas in the same way.

This finding sheds light on one of the controversial issues in the literature on impact evaluation, that is, the role of Humanities and Arts. It seems that one of the tenets of the argument that Humanities and Arts are not sensitive to the audiences, or users, of their research results, is simply false. When asked to reflexively reconstruct their impact, they are systematically able to mention their target groups.

This comes at a cost, however. Humanities rank top in the Frequency and Diversity of target groups, but rank at the bottom when coming to their Specificity. Figure 13.7 shows an almost reverse ranking of subject areas when the indicator chosen is Specificity. At the top of the Specificity ranking we find Social Sciences, such as Economics, Politics, Anthropology and Law. This is another interesting finding. These disciplines have been able to target highly specific groups of users, following a process of academic specialization. They have low Frequency and low Diversity- that is, do not spend much emphasis on their target users, but have high Specificity- that is, they know whom to target.

Large part of Humanities are found at the bottom: Psychology, Education, Philosophy, Language and literature, Art and design. Engineering disciplines are more scattered.

Finally, by combining the three indicators it is possible to examine more closely the pattern of impact of subject areas.

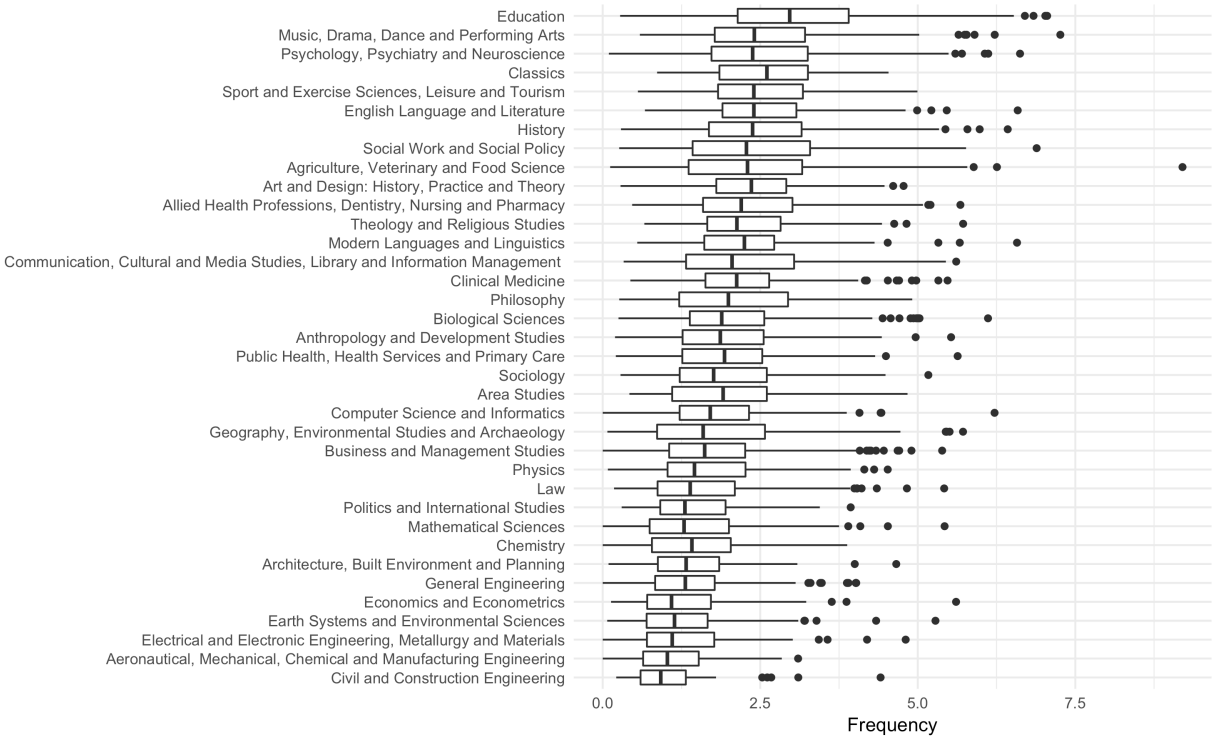

Figure 13.5: Boxplot of the Frequency indicator in the corpus of REF impact case studies by subject area (Unit of assessment)

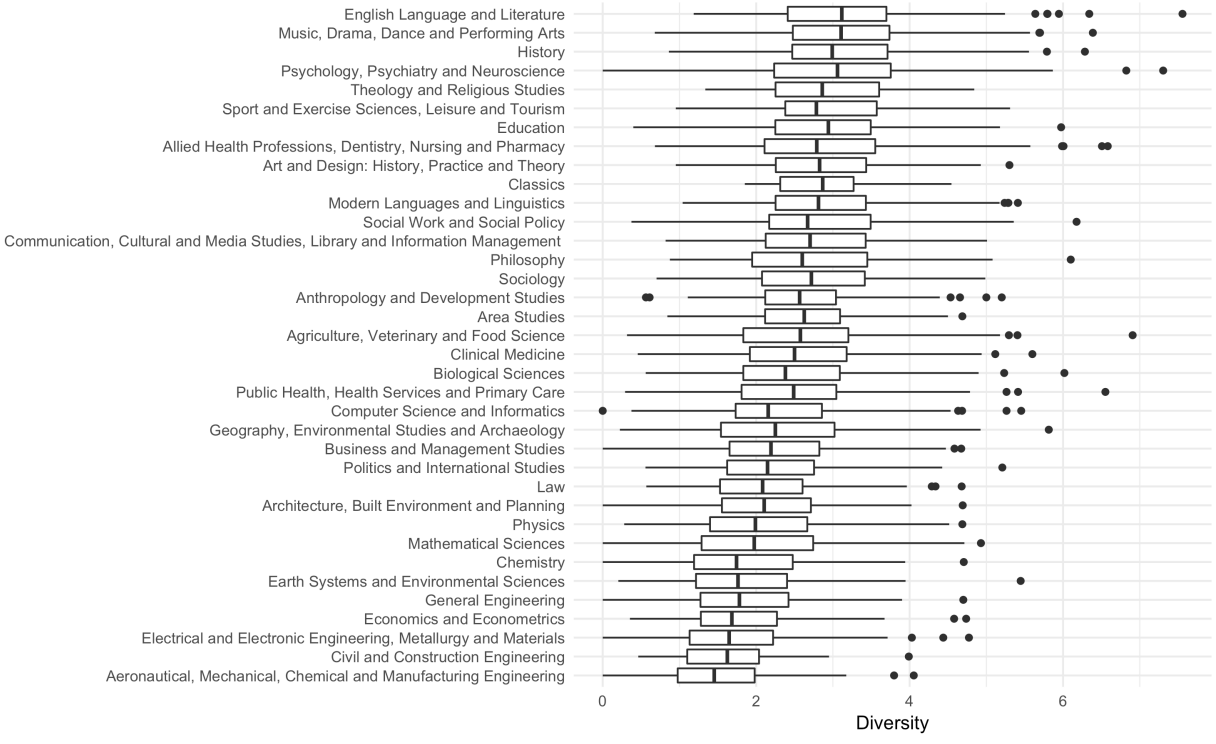

Figure 13.6: Boxplot of the Diversity indicator in the corpus of REF impact case studies by subject area (Unit of assessment)

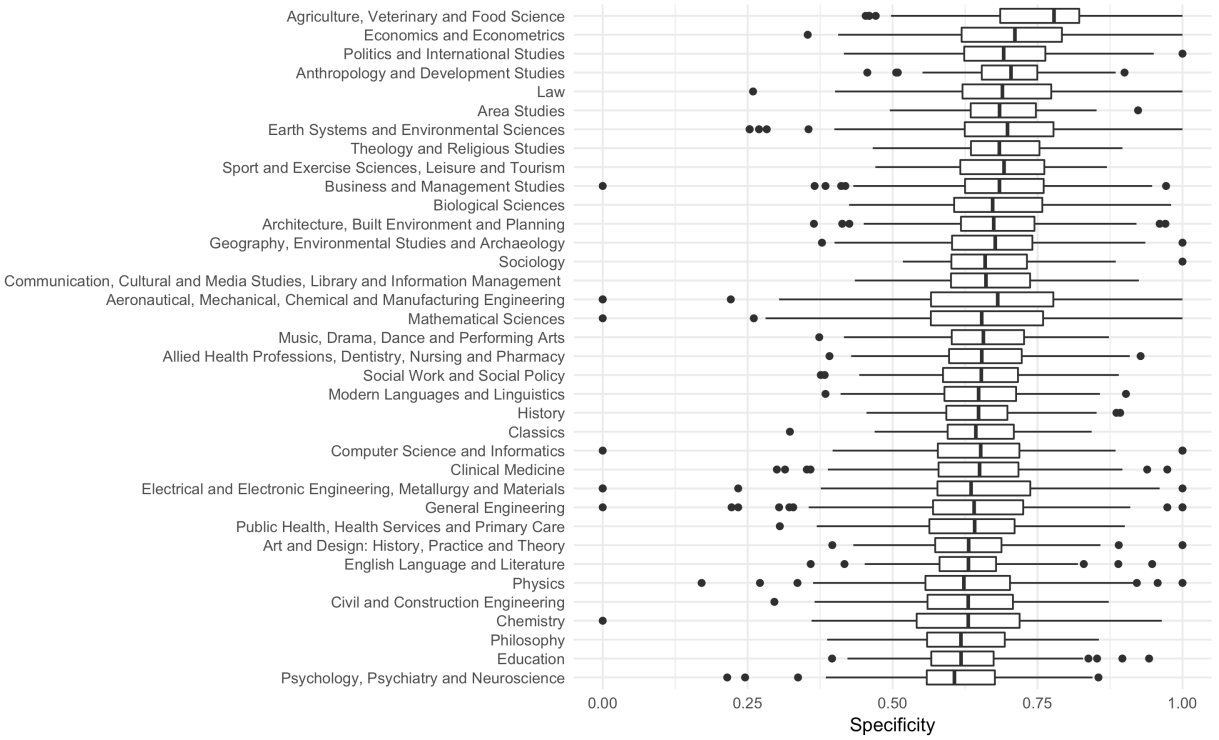

Figure 13.7: Boxplot of the Specificiy indicator in the corpus of REF impact case studies by subject area (Unit of assessment)

Frequency and Diversity are highly correlated (13.8). Humanities rank top in both indicators and are located in the top right region of the graph. Almost all disciplines in Engineering and Economics and Econometrics lie at the opposite corner. When coming to Specificity, correlation with the other indicators is on the contrary extremely low and is negative (- 0.016 with Intensity, - 0.146 with Diversity).

13.6 Discussion

The data show an intriguing disciplinary pattern: Humanities and Arts show remarkably higher frequency of terms related to users, but a significantly lower specificity. The opposite is found for many areas of STEM, namely Engineering and several Natural science disciplines.

The results for Humanities are very intriguing. Research in Humanities is often considered as pure, abstract, not engaged with society. This representation is used by those governments that argue that research funding in these areas should be cut because their impact on society cannot be demonstrated. For Social Sciences the situation is slightly different, but there is a presumption that only a few social sciences have an impact, in particular those with instrumental value, such as Economics. Our data tell a different story. When asked to demonstrate the impact of their research, scholars in Humanities and Arts use a very rich vocabulary of users and mention users very frequently, twice as much as scholars in STEM. At the same time, they are less capable to transform their orientation in operational terms, by definining, identifying, targeting specific groups or audiences. At the opposite extreme, it seems that scholars in STEM mention very precise and well defined groups of users. This might be a result of the nature of their studies: researchers in Medicine follow specific targets in terms of disease and patients, or technologists have narrow industrial applications. This remarkable difference may create an imbalance in the assessment of research impact.

Research in STEM finds more easily and unambiguously the groups of potential users, so that its impact

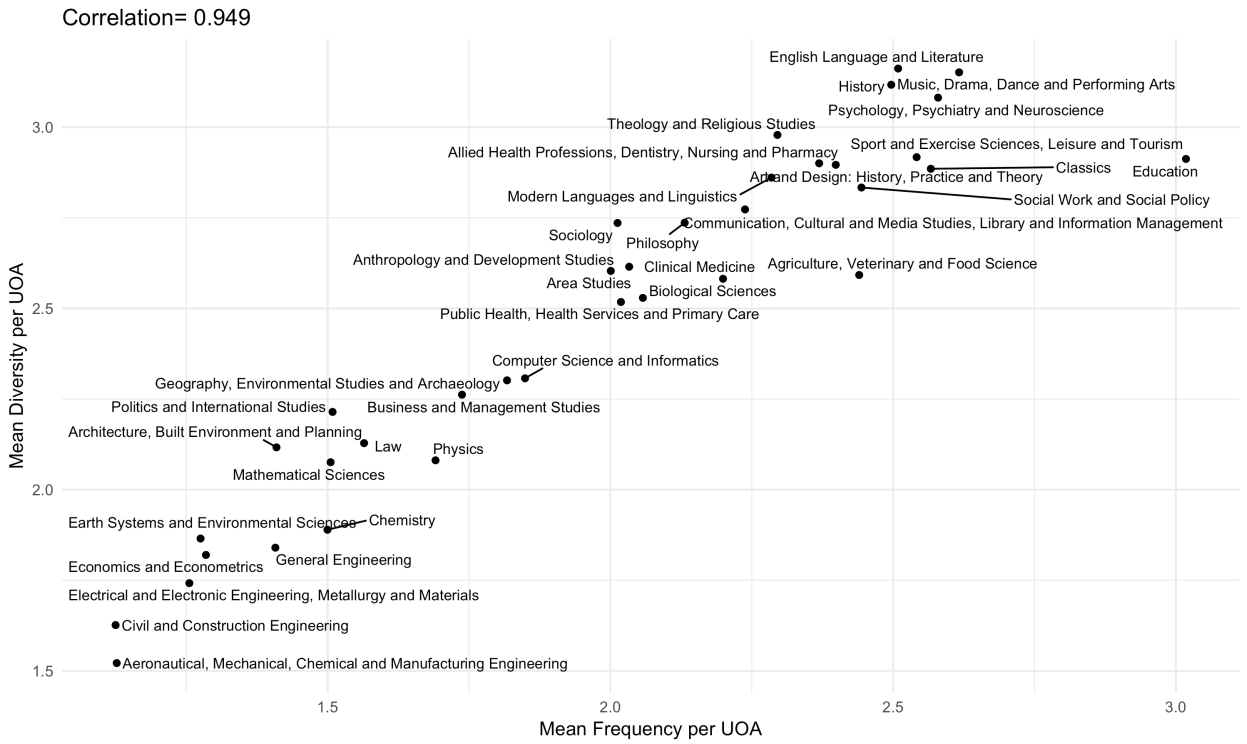

Figure 13.8: Scatterplot of the relation between Frequency and Diversity indicators in the corpus of REF impact case studies by subject area (Unit of assessment)

is easier to observe and operationalize. As it has been noted by the studies on REF, evaluators may find it easier to use concepts drawn from STEM to evaluate all types of research, simply because they point to more observable and discrete products of research. From the field of social studies of evaluation it is well known that measurements lead to a feedback loop so that the immediate availability of indicators may lead to believe that important things are only those that can be measured.

The notion of impact of research deserves further and intense research effort in the near future. We see several directions of research. First, apply the methodology to ex ante research documents, such as research proposals: do they include users? do they identify target groups with adequate Diversity and/or Specificity? Second, test whether there is a relation between the Specificity of identification of target groups and the assessment of the research, either ex ante (approval of a proposal) and ex post (score obtained in research assessment exercises). Third, extend the methodology to other kinds of documents.

Chapter 14

Defining Industry 4.0 Professional Archetypes

Worldwide industrial systems are evolving by leveraging Internet connected technologies to generate new added values for organizations and society. Researchers, policy makers and entrepreneurs refer to such phenomenon with the name of Industry 4.0.

An increasing number experts from different fields are focusing on this topic, bringing their contribution in terms of new technologies and methods. As a consequence of this process, the companies that are embracing the new paradigm need to manage new technologies and the new relations among them with a multidisciplinary approach. The result is an emerging need of new professional figures able to bridge the different fields.

Moreover, while the scientific interest in technological aspects of Industry 4.0 is constantly growing, the understanding of the future works and professional roles is slowed down by the heterogeneity, complexity and static nature of job description systems. These issues are usually addressed by qualitative methods thus making the results uncertain and partial.

The first step of the present chapter is to develop a data driven mapping of 4.0 competencies which benefit (rather than being disadvantaged) by the heterogeneity of the entities to map. Then we propose a classification of the groups of competencies with the aim of identifying and defining the archetypes of Industry 4.0 workers. Given the bottom-up process we adopted to reach our goal, also the relationships between them can be described.

14.1 Digital Competences Development

The radical technological change is going to affect the whole industrial environment and will radically transform, in several different ways, the world we live in (see section 9.1). Consequently, the revolution is not just referred to the availability of new sophisticated machines, but also to the deep reconsideration of worker roles it brings with. On the one hand Industry 4.0 implementation may have negative impact on specific professional figures, which could be substituted by robots. In support of this theory, Osborne & Frey implemented a methodology to estimate the probability of job computerization using a Gaussian classifier (Frey and Osborne, 2017). The results are quite pessimist: they distinguished between High, Medium and Low risk of computerization: according to their estimates around the 47% of jobs is in the high risk category. On the other hand, another possible impact is represented by a significant increase of worker competences and the emersion of new professional profiles: our research mostly focused on the second prospect.

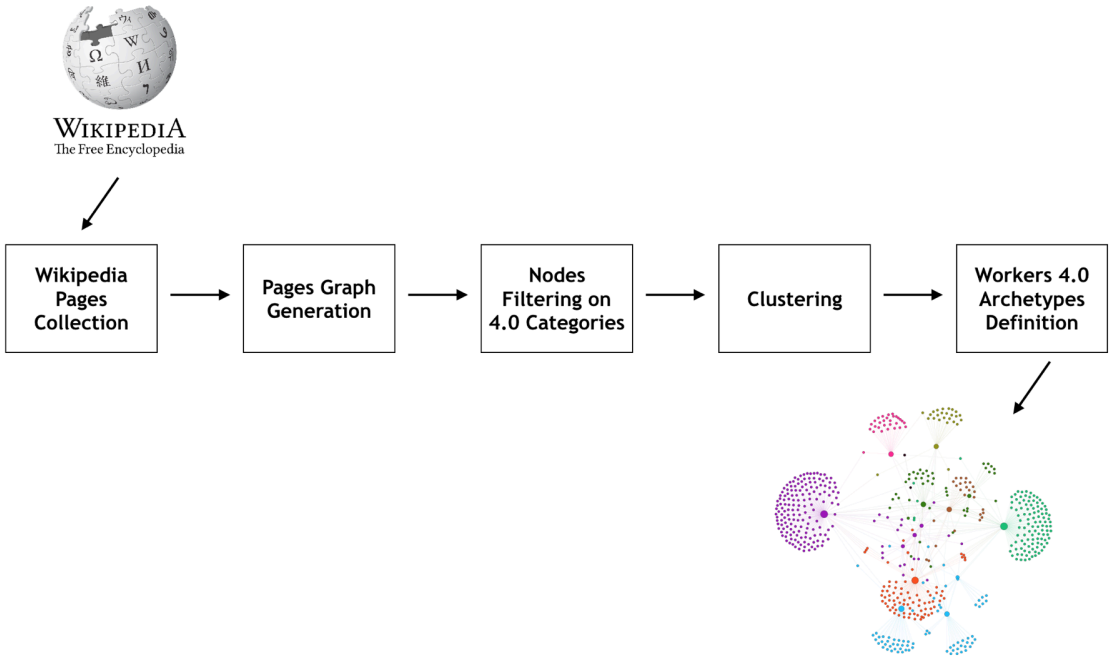

Figure 14.1: Flow diagram of the process of 4.0 Archetypes definition.

14.2 Methodology

In this chapter we show the process to built the graph of Wikipedia pages related to Industry 4.0. This graph will be then used by expert to define 4.0 workers archetypes. Figure 1 shows the process we adopted. The whole process takes in input a set of linked entities (e.g. Wikipedia pages) and gives as output a clustered graph.

Wikipedia Pages Collection

The process takes as input a set of entities linked between them. A link between two entities represent the existence of a relation between them. Here the concept of relation is intended as the sharing of a topic. In other words, two entities has to be connected if they belong to the same topic. Any entities-links structure that meets this requirements can be used as input for the proposed process. For the purpose of the present work, we decided to use the free encyclopedia Wikipedia . Wikipedia is structured in such a way that each page contains links to other Wikipedia pages. The links from page to page are manually assigned by the contributors, so the structure of the links evolves dynamically. Furthermore each page is labeled by the contributors through categories with the intent of grouping together pages on similar subjects. To collect industry 4.0 related wikipedia pages we exploit this information and structure. In the date of 23/11/17 we collected three levels of pages typologies for a total of 4739 pages:

- L1, The Wikipedia page of industry 4.0 [1] *(1 page)*
- L2, All the pages linked to L1 *(39 pages)*
- L3, All the pages linked to L2 *(4699 pages)*

For each page we collected the page name, the links and the categories.

[1] https://en.wikipedia.org/wiki/Industry_4.0

Pages Graph Generation

Once we collected all the pages and the links between them, we represent this structure as a directed graph (S). It is composed of:

- A set of nodes (N): the Wikipedia pages.
- A set of edges (E): the links between the pages (considering also the direction of the edge)

S has 4739 nodes and 194.299 edges.

Nodes filtering on 4.0 Categories

Considering the content of Wikipedia, it is reasonable to assume that not all of the 4.739 pages are related to Industry 4.0.

To sanitize the set of nodes N we adopted a series of filtering rules based on the categories of the pages. The steps we followed were:

- Count the occurrence of each category we extracted. At this step we had 14.711 categories.
- Filter the categories having an occurrence minor or equal to 3. This threshold level has been chosen looking at the distribution of the occurrences (the number of pages per category). A change in the central node could make changes on the selection of this threshold level. At this step we had 1.605 categories.
- Manually screen the categories and select those related to industry 4.0, taken in to consideration the definition of Industry 4.0. At this step we had 337 categories.
- Selecting the nodes that contains at least one of the 337 categories.

As an example, the top 10 categories in terms of occurrence and the relative occurrences are: production and manufacturing (51), parallel computing (42), manufacturing (37), business terms (33), management (31), process management (29), engineering disciplines (28), internet of things(28), technology in society (26) and cloud computing (25). This list shows how the process is able to make categories correlated to Industry 4.0 emerge. We thus obtain a new graph S' that is a subgraph of S (all the nodes N' and the edges E' are a subset of N and E respectively). S' has 645 nodes and 703 edges. Of these nodes 75 are disconnected from the graph (both in-degree and out-degree equals to zero) and are thus not useful for the analysis we will conduct in section 3.4 . These nodes are filtered, obtaining a definitive S' having 570 nodes.

Clustering

We can assume that S' gives us a reasonable representation of Industry 4.0 related wikipedia pages in terms of precision and recall.

It is thus possible now to analyze if the pages are arranged in clusters, or in other words, if there exists groups of similar pages.

To understand the concept of similarity we need to consider the nodes N1 and N2, and the sets of all nodes connected to N1 and N2, respectively L1 and L2. The similarity of N1 and N2 is proportional to the module of the intersection between L1 and L2. To investigate this similarity and to find clusters of similar pages, we grouped similar nodes via short random walks. The algorithm is described in [Durrett, R.: "Random Walks. Random Graph Dynamics", 153-186.]. The approach finds ways to compute an intimacy relation between the nodes incident to each of the graph's edges. The process is non-supervised so it makes an optimal decision on the number of clusters. The output was 9 different communities.

The obtained graph is shown in Figure 14.2. The figure shows the nodes, the edges and the label of the nodes. The dimension of the nodes is proportional to the out-degree of each one (the number of tail edges adjacent to the vertex) and the color represent the community to which it belongs to. The label of the nodes (as center of the clusters) are shown for those having an out degree greater than 4. The table shows the correspondence between the community, the colors and the percentage of nodes per community.

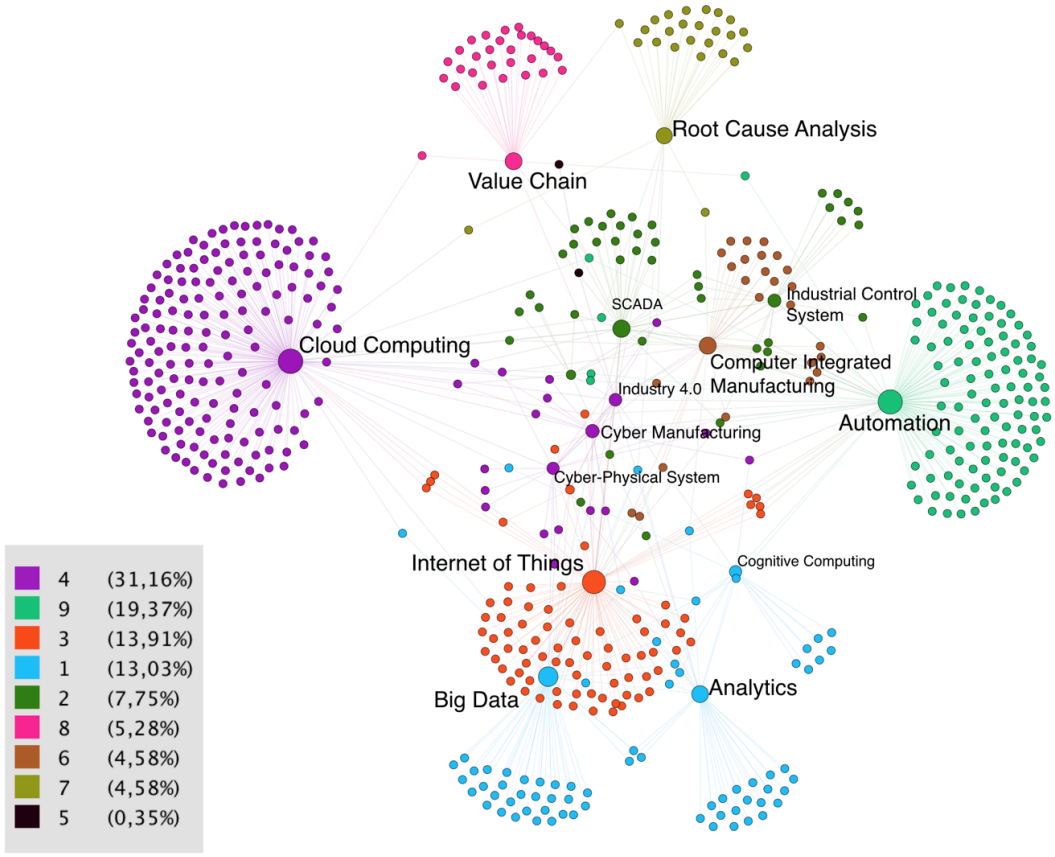

Figure 14.2: Representation of the graph S'. The nodes are the Wikipedia pages related to Industry 4.0. The arch between two nodes exists if there is a link between the pages. The labels of the nodes are shown only for nodes having an out-degree (number of tail edges adjacent to the vertex) greater than 4. The color of each node represent the community to which the node belongs. The table in the low-left shows the percentages of nodes for each community.

		Community Number									
		1	2	3	4	5	6	7	8	9	Total
Archetypes Names	the Architect	44	1	3	6	0	4	0	0	5	63
	the Prophet	21	0	4	0	0	0	0	0	7	32
	the Perfectionist	0	36	1	2	0	0	0	0	3	42
	the Geek	5	1	65	152	0	10	0	0	43	276
	the Investigator	0	0	0	0	0	2	26	0	2	30
	the Strategist	0	0	2	4	0	6	0	30	9	51
	Null	5	6	4	14	2	4	0	0	41	76
	Total	75	44	79	178	2	26	26	30	110	570

Figure 14.3: Table of the accordance between the archetypes and communities. The total of pages for each archetypes is shown in the last column, the total of pages for each community is shown in the last row.

Workers 4.0 Archetypes Definition

The authors generated a list representing the "Industry 4.0 workers archetypes" based on Wikipedia. The bias due to the profile of people writing this kind of Wikipedia pages does not reduce the importance of this exercise, but it is fundamental to understand why the reader will find such a great emphasis on technologies, methodologies and tools as pages describing the archetypes in section 4. Skills, competences and professional figures (in the strict sense) are then behind and beyond the archetypes, which represent just the attitudinal profile of the workers that will be soon seek by companies adopting Industry 4.0 paradigm.

Table of Figure 14.3 shows the relation between the Archetypes and the communities (created by the text-mining procedure presented above). It also shows the total number of nodes per archetypes and per community.

From this table it emerges that:

- Group 1 (75 pages) has two nodes/centers: the page "Big Data" and the page "Analytics". This is why from Group 1 derives two different Archetypes: "the Architect" (44 pages, centered in Big Data) and "the Prophet" (21, centered in Analytics). This is an evidence of the fact that these two archetypes are strongly related between them, since they belong differently from all the other groups.
- More than 80% of the pages referred to Group 2 (centered in the page "SCADA" and "Industrial Control system) were used for creating the Archetype of "the Perfectionist". It is important to underline that the page SCADA - acronym of Supervisory control and data acquisition- being a specific control system architecture, was discarded from the assignation to a specific archetypes. Our focus for designing "the Perfectionist" was the generic page "Industrial Control System".

- Group 3, Group 4 and Group 9 are composed of pages we linked to the same Archetype: "The Geek". In respect of Group 1, we made a different choice: we referred the most part of these three groups' pages to just one Archetype, "nuanced" on the basis of the nearest technologies in the graph: "Cloud Computing", "Automation" and "Internet of Things". As the Geek was created starting from three different groups, it has three different declinations as shown in Figure 14.4. The geek, as will be explained in section 4, has to be seen as the expert of a specific family of similar technologies strongly related to industry 4.0.
- Group 6 is composed of 26 pages, that were referred to almost all the Archetypes. These pages are general purpose, and referring to Figure 14.2, this is also evident from their central position in the graph (there are related to every community). The predominance of pages referred to one Archetype (the Geek) didn't allow us to consider this Group, centered in the page "Computer Integrated Manufacturing", as another nuance of "the Geek". Moreover, Group 6 is strictly related to Group 9, as the proximity of the clusters in the graph can easily show.
- In group 7, centered in "Root Cause Analysis" page, we found 100% of pages referable to the Archetype we called "The Investigator".
- As in the previous group, also in Group 8 (centered in "Value Chain" page) the 100% of pages were referable to one Archetype: "the Strategist"

Figure 14.4 shows the same graph of Figure 14.2, associating to each group an Archetype. In this new version of the graph it is possible to identify the centers of each cluster, the position of the clusters (and the related archetype) in respect of the others, the interrelation among the clusters/archetypes.

14.3 The Archetypes

The term "archetype" has its origins in the Greek word "archein", which means "original or old" and "typos" "pattern, model or type". It is thus etymologically referred to the "original pattern" from which all other similar persons are derived, modeled or emulated. In this work we consider this etymological meaning, discarding any philosophical or psychoanalytic references which would mislead our research. As regards the personality many theories have been proposed over the years [11,12], but a common vision at the academic level has not been reached yet. Both Jung's theory of personality types [13] and the Myers-Briggs indicator (MBTI) in the field of work psychology [14] come to several psychological types defined through descriptions and traits. Some of them (e.g. the 16 psychological types of MBTI) are a standard for Human Resource managers and are largely used for resource classification and selection. It is important to underline that in the present work i) technologies and skills have been extracted in a scientific and repeatable manner, ii) they have been clustered following a well-known and accepted algorithm. Conversely, archetypes' names have been chosen with the intent to better communicate the research outcomes.

The archetype description sections are built following a common structure:

1. *The name of the archetype.* The choice of the name was made by the authors in order to have an evocative impact on the reader and reaching a communication goal. In some cases, the names are deliberatively imaginative, renouncing to respect a rigid scientific approach with the aim of charming and interesting the reader
2. *The list of the Wikipedia pages selected.* The pages represent the elements considered by the authors for defining the archetypes. The selection of the pages was done referring to the clusters identified with the data-mining as shown in section 3. All the Archetypes are referred at least to one group, with the exception of group 1, from which two different Archetypes were created, and groups 3,4 and 9, which all together participated in the creation of another Archetype (even if with three different declinations). In many cases the pages reported in the box are related to technologies, methodologies, tools: evidently the authors of Wikipedia pages related to Industry 4.0 are more interested in "hard" rather than "soft" topics, such as competencies or cultural aspects of the new paradigm.

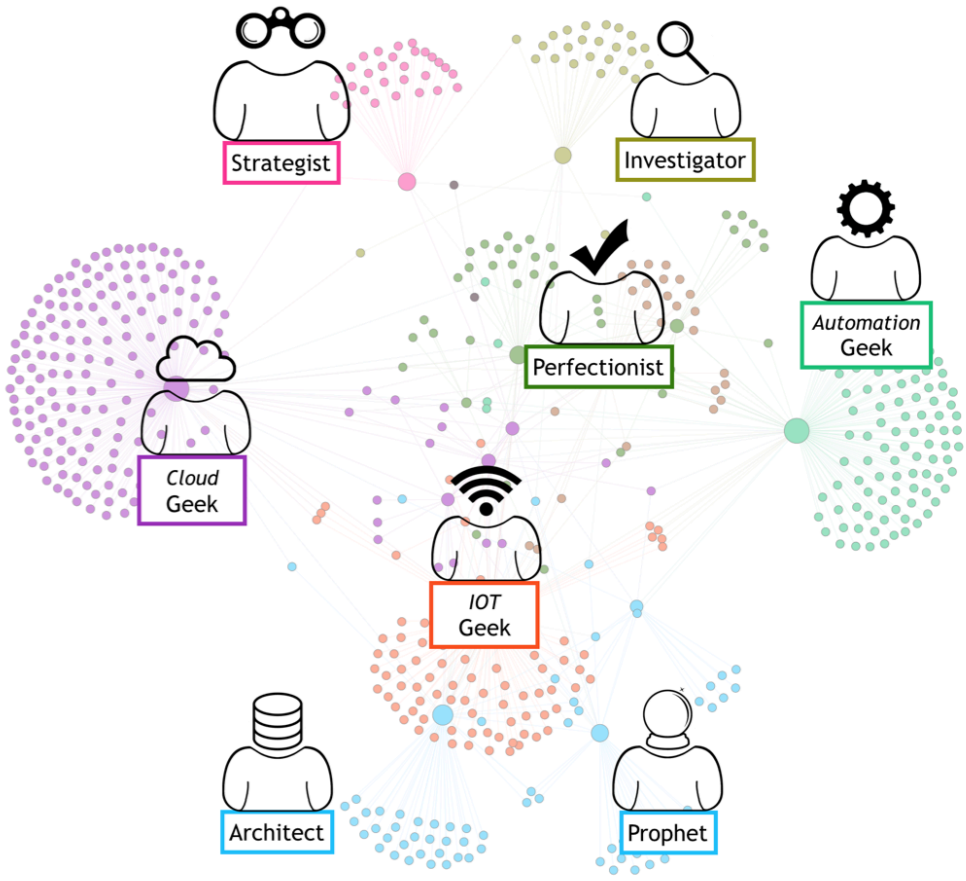

Figure 14.4: The different professional archetypes associated with each technological cluster.

3. *Industry 4.0 Emerging needs.* The authors presented shortly the needs that the companies face when introducing Industry 4.0 applications. The needs are mainly based upon the interpretation given by the acatech study [15], which represent one of the most effective key for understanding the new paradigm [16].
4. *The archetype.* We have already explained what we meant with the word "archetype" and the rationale behind the choice of the specific names. The explanation of the archetype given in this section is based on the interpretation of the Wikipedia pages selected and on the semantic fields they refer to. We tried to describe the inner nature of the individuals referable to each specific archetype, defining the peculiar traits of their mindset and their most marked features. The aim was also to describe them so to make them recognizable in our daily experience.
5. *Keywords.* The keywords are taken from the column "categories" of the clustered list of selected Wikipedia pages as shown in section 3.2. Strikethrough in the body of the text indicates where the authors decided of deleting misleading words: we kept them for making clear and transparent the procedure behind this cleaning activity. The reasons for deleting some words were: a) they were too generic and did not characterize the archetype, b) they were too specific and they made the archetype losing its capacity of encompassing more aspects, c) even if they survived to the "cleaning" after the data-mining, they were referred to another semantic field (disambiguation).

The Architect

The Wikipedia pages selected: Big data; Data blending; Data quality; Cognitive computing; Data fusion; Data science; Unstructured data; Data lake; Data set; Data; Data lineage; Data transmission; Radio-frequency identification; Data philanthropy; Datafication

Industry 4.0 Emerging needs

When taking a decision in Industry 4.0 paradigm it is more important than ever that the individuals can access, collect and process data and information. Data can be gathered from machinery, equipment and tools thanks to sensors, actuators and information processing systems, linked to a communication layer. Big data is a buzzword usually linked to Industry 4.0 and used to describe mass data that cannot be analyzed adopting common procedures, but requires advanced applications and technologies. Thus, in the new paradigm, it is important not just collecting the data, but also collecting and processing them in an innovative way and shaping them in a form that make them usable for the company decision-makers and other employees.

The archetype: The Architect

The Architect is the individual that is at ease in managing, transforming and processing data. He/she has a clear idea of the importance of data and understand which data are useful, in which context and for who. The Architect likes formalized information and the possibility of shaping them, has an innate capacity of breaking the problems and find the best solutions: this capacity is fundamental when having to deal with different sources of information and numerous data and inputs. He/she is precise and reliable, has a great aptitude for visualization and for problem solving.

Key words: big data; data management; distributed computing problems; technology forecasting; transaction processing, artificial intelligence; cognitive science business intelligence; data; information technology management, computer data, automatic identification and data capture; privacy; radio-frequency identification; ubiquitous computing; wireless, big data, types of analytics, online analytical processing

The Prophet

The Wikipedia pages selected: Analytics; User behavior analytics; Business analytics; Data mining; Data analysis; Continuous analytics; Machine learning; Data visualization; Cultural analytics; Predictive analytics; Natural language processing; Customer analytics; Business intelligence; Analytic applications; News analytics; Statistics; Behavioral analytics; Predictive engineering analytics

Industry 4.0 Emerging needs

The implementation process of the Industry 4.0 paradigm involves different stages. At one advanced stage of this process the company should be able, starting from the data collected, to simulate future scenarios and select the most likely one [15]. To do so the company needs to have analyzed and assessed a number of data and projected the "digital shadow" of its assets in the future. Forecasting and the recommendations based on it became the key elements on which building the development of the company. The analysis of the different scenarios is the basis for building the business success. For reaching these targets it is fundamental to have constructed proper "digital shadows" of the assets and have understood the interactions among them.

The archetype: The Prophet

The Prophet is an individual with strong analytical thinking skills, able to examine in details specific situations in a critical way. He/she has a future-oriented mindset: is at ease in thinking about next steps, possible scenarios and with a medium-to-long-term vision. Starting from the analysis of data and adopting also statistical methods, the Prophet investigate the problems (creative thinking) and identify possible solutions (complex problem solving) in order to reach the best case scenario he designed.

Key words: analytics; business intelligence; business terms; financial data analysis; formal sciences; data mining; formal sciences; cybernetics; learning; machine learning; actuarial science; big data; business intelligence; financial crime prevention; prediction; statistical analysis; types of analytics; financial technology; information management; data; formal sciences; information; mathematical and quantitative methods (economics); research methods; statistics; artificial intelligence; computer security; human behavior; machine learning; computational fields of study; data analysis; particle physics; scientific method; data; information technology governance; statistical charts and diagrams; visualization (graphic); artificial intelligence; computational fields of study; computational linguistics; natural language processing; speech recognition; business intelligence; business software stubs; applied data mining; business analytics; market research; marketing performance measurement

The Perfectionist

The Wikipedia pages selected: Industrial control system; Instrumentation; Control valve; Overall equipment effectiveness; Enterprise resource planning; Distributed control system; Control system; Lean manufacturing; Total productive maintenance; Programmable logic controller; Risk; Control System Security; Control loop

Industry 4.0 Emerging needs

Since the First Industrial Revolution, quality improvement has been one of the critical aspects for developing any business. In the current age of high competition and mass production, this has become, along with price strategies, the main element for expanding the market. Thus, quality control has become a central issue to consider before establishing any industrial undertaking and the best way for ensuring the best allocation of resources and highest level of production. In Industry 4.0 paradigm quality has become even more important, mainly because the increased flexibility of machinery, processes and procedures allow to intervene continuously for analyzing the parameters, identify machine faults or quality issues with a high degree of confidence and in advance. Flexibility and agility (which denotes the ability to implement changes in real-time, including changes to the company's business model) are two key factors that allow a constant improvement.

The archetype: the Perfectionist

The Perfectionist is the individual who is never satisfied of the actual conditions: it could be his/her current personal situation, the preparation for an exam, the functioning of a process or a machinery. He/she always thinks that things can be done better, in a shorter time, with less resources, involving different teams, collaborating more (or less). The Perfectionist likes tests: he/she tries, monitors, checks the results, thinks about it (possibly analyzing data and empirical results) and then suggests - or decides - how to make things work right.

Key words: control engineering; industrial automation; industry; manufacturing; telemetry; lean manufacturing; automation; control theory; systems engineering; systems theory; industrial computing; programmable logic controllers; measuring instruments; sensors; business terms; computer-aided engineering; erp software;

enterprise resource planning terminology; information technology management; management; production and manufacturing; supply chain management terms; commercial item transport and distribution; inventory; management; process management; production and manufacturing; actuarial science; environmental social science concepts; financial risk; risk; applications of distributed computing; control engineering; industrial automation; control devices; valves; applications of distributed computing; control engineering; industrial automation; business terms; commercial item transport and distribution; computer security; computer security procedures

The Geek

The Wikipedia pages selected: Cloud computing; Smart grid; Productivity; Cyber manufacturing; Artificial intelligence; Embedded system; Cyber-physical system; Cybernetics; system; Manufacturing; Computer network; Cloud manufacturing; Computer security; Internet of things; Information technology; Distributed computing; Service-oriented architecture; Automation; Adaptable robotics; Automation Technician; Developmental robotics; Industrial Engineering

Industry 4.0 Emerging needs

Beside the aspects related to a new organizational structure and a different work culture, Industry 4.0 is strongly determined by the new technologies, both hardware and software. It is impossible to implement Industry 4.0 applications without considering technical aspects. The enabling technologies (Clouds, Augmented reality, Simulation, IoT, etc.) must be known in the company in order to make the new paradigm real. Technologies are so specific and so many, that it is important to have an extensive knowledge of them (even if not deep) and be able to integrate them for finding the most effective custom-made solution to solve specific problems.

The archetype: the Geek

The Geek is the archetype of an individual extremely passionate in new technologies and its applications. He/she is interested in everything is new and innovative, in seeing beyond and within things: how technological applications work, why and at what extend their capacities can be brought. Integration is always an important aspect: the Geek appreciate the ideas of combining, breaking and keeping just some aspects for recombining and then creating something new. He/she has a kinesthetic learning (or tactile learning) and has a kinesthetic approach to the work: he/she needs make tests (rather than thinking or planning), piloting solutions (rather than have someone else doing the job) and working on the field. The geek is always up-to-date on the newest technology and is always willing to be the first in creating something new. As the number of pages referable to the Geek were numerous, the authors decided to create three different nuances represented by three different logos, one for each field of expertise of the Geeks.

Key words: ambient intelligence; emerging technologies; internet of things; computers; information-theoretically secure algorithms; information technology; media technology decentralization; distributed computing; architectural pattern (computer science); enterprise application integration; service-oriented (business computing); software design patterns; web services; cloud computing; cloud infrastructure; big data; industrial revolution; industrial automation; industrial computing; technology forecasting; computer systems; industrial computing; industry; internet companies; manufacturing; computer systems; physical systems industry; manufacturing ambient intelligence; emerging technologies; smart grid; artificial intelligence; computational fields of study; computational neuroscience; cybernetics; formal sciences; technology in society; unsolved problems in computer science; cybernetics computer networking; computer networks; telecommunications engineering; computer network security; crime prevention; cryptography; cybercrime; cyberwarfare; e-commerce; information governance; national security; secure communication; security technology; weapons countermeasures; economic growth; industry; production and manufacturing; production economics; electrical engineering; embedded systems; cloud computing; automation; engineering disciplines; adaptable robotics; robot kits; robotics stubs; industrial automation; technicians; machine learning; robot control; engineering disciplines; industrial engineering; operations research; production and manufacturing

The Investigator

The Wikipedia pages selected: Root cause analysis; A3 problem solving; Failure mode and effects analysis; Quality control; Business process mapping; Forensic engineering; Business process; DMAIC; Incident management; Value stream mapping; Design of experiments; Pareto chart; 5S (methodology)

Industry 4.0 Emerging needs

Industry 4.0 paradigm enables to a constant check on procedures, processes and systems, as they are all integrated and all the information are conveyed to a same "repository" (the communication layer). All these aspects can be then monitored in a totally new way: it is possible, for example, to have data in real time and the decisions can be taken according to information which can be transferred almost at the same time they are created by one or more devices or tools. Thanks to the high amount of information than can be analyzed, also mistakes can be treated in a different way: in Industry 4.0 it becomes easier to focus and investigate on the causes of a failure – and solve it -, rather than on finding out who is to blame.

The archetype: the Investigator

The Investigator is the individual who is naturally curious: as a proper scout he/she wants to understand what is working and what is not. He/she is never satisfied with the explanation he/she receives and is always searching for a mistake, a bug or just another way for reaching at the same conclusion. The Investigator is eager for analyzing what is already known and what is new, he/she likes going into details and understanding how things works and how they could work in a different way. When identified the problem the Investigator is not necessarily interested in finding the fix, he/she rather prefer to have some other problems to break and leaving someone else the possibility of coming up with a solution (that he/she would love to analyze again).

Key word: business terms; problem solving; process management; production and manufacturing; quality; quality management; six sigma; design for x; management; quality control; statistical process control; business process; enterprise modelling; process management; lean manufacturing; commercial item transport and distribution; inventory; japanese business terms; lean manufacturing; manufacturing; methodology; process management; quality control tools; business process; data management; design of experiments; experiments; industrial engineering; quantitative research; statistical theory; systems engineering; reliability analysis; reliability engineering; systems analysis; engineering disciplines; engineering failures; forensic disciplines; materials science; business software; disaster preparedness; enterprise modelling; firefighting in the united states; itil; incident management; categorical data; statistical charts and diagrams

The Strategist

The Wikipedia pages selected: Value chain; Core competency; SWOT analysis; Business model; Delta Model; Strategic management; Balanced scorecard; Game theory; Strategic planning; Control (management); Porter's five forces analysis; Strategic thinking

Industry 4.0 Emerging needs

If Industry 4.0 allows the companies to foresee different scenario based on data and information, it doesn't not guarantee to the entrepreneurs to select the right one or take the correct decision. Data should be read correctly, but also the strategy for reaching the preferred scenario should be decided in the proper way. In particular, one of the most dangerous risk is not being able to redefine properly the company business model, which should be constantly monitored, assessed and, according to the most updated information, modified.

The archetype: the Strategist

The Strategist is the individual who has a mind wide open: he/she thinks to the outcomes of each activity and reflect on the different ways for reaching the target. Strategic thinking, strategic management and planning are its most marked qualities: he likes thinking to all the possible outputs in a comparative way. He/she is also a pragmatist, as starting from the analysis of the current situation he/she doesn't like to create different scenarios (as the Prophet does) but rather reaching the result established in the more effective way. Even if the Strategist is not an entrepreneur, he has some of its qualities and this is why this archetype can

be referred to the "entrepreneurial mindset". The Strategist is visionary, risk-taker, creative, with a strong willingness to challenge the status quo and he has diplomatic and leadership skills.

Key words: distribution (business); michael porter; process management; supply chain management; value proposition; business models; management; strategic management; types of marketing; business planning; business software; business terms; management; control (social and political); control theory; artificial intelligence; formal sciences; game theory; john von neumann; mathematical economics; business terms; strategic management; business intelligence; strategy; systems thinking

14.4 Conclusion

The present section can be seen as a first step of a long journey. As stated in introduction our Archetypes are intended in the etymological sense, as original models, and we assume that they are emerging in the digital economy. Thanks to the data-mining exercise we were able to place the clusters of Wikipedia pages in a graph, analyzing them according to the contents, the position in the graph and the relation among them. This recurrence of same pages in almost all the clusters let us concluding that some topics are common to all the Archetypes, independently to the cluster they were originated. For example, being able to read and exploit data is a transversal issue and all the Archetypes (and all the workers referable to them) should be used to work with them. Another constant in all the cluster is the reference to technologies: they are central in the new paradigm, it is unavoidable for a digital worker knowing them. Even if the "Fourth Industrial Revolution" doesn't end with a bunch of IT tools, they are at the center of the new paradigm and workers must be used to play with them [17].

To conclude and anticipate the further developments, in figure 14.5 we give an interpretation of the position of the archetypes in the graph. We can cluster the Archetypes according to the general company area they could be referred. The areas are:

1. Business. The Strategist and the Investigator are the main actors. In this area the decision regarding the future and the strategy of the company are taken: it is important to have people who have always the big picture clear in their mind (the Strategists) and are good in analyzing the situation for what it is (the Investigator).
2. Data. The Architect and the Prophet suits well here. This area is referred to the acquisition, collection and management of data. This area is perfectly suitable for someone who is comfortable in analyzing and structuring data (the Architect) and for who has a future-oriented mindset, always based on the actual inputs coming from the company functions (the Prophet).
3. Process. The Geek (automation) is the only major player in this area. The process area is where the core activities of a company are performed.

The connection between the three areas are ensured by three Archetypes:

- The Perfectionist, is the bridge between the process area and the business area: in both of them is useful to have someone who is comfortable in solving problems and making things work always better.
- The Geek (cloud) is the bridge among the Data area, where he/she contribute with the knowledge of technologies, and the Business area, where Archetypes will benefits from these technologies for new business development.
- Finally, the Geek (IoT) is the connection among the Process area and the Data area, representing the bridge between the core activities and the usage of information coming from them.

A possible next step of our study will be identifying at which function of the Porter value chain model each Archetype suits better. Then, having placed the Archetypes in their most suitable function, it will be then possible to allocate for each Archetype the most appropriate 4.0 professional profiles, so to follow their attitudes and exploit better their skills. At this stage, as the new paradigm is still emerging, we cannot go further in our research, but the road-map is now clear and the first step is taken.

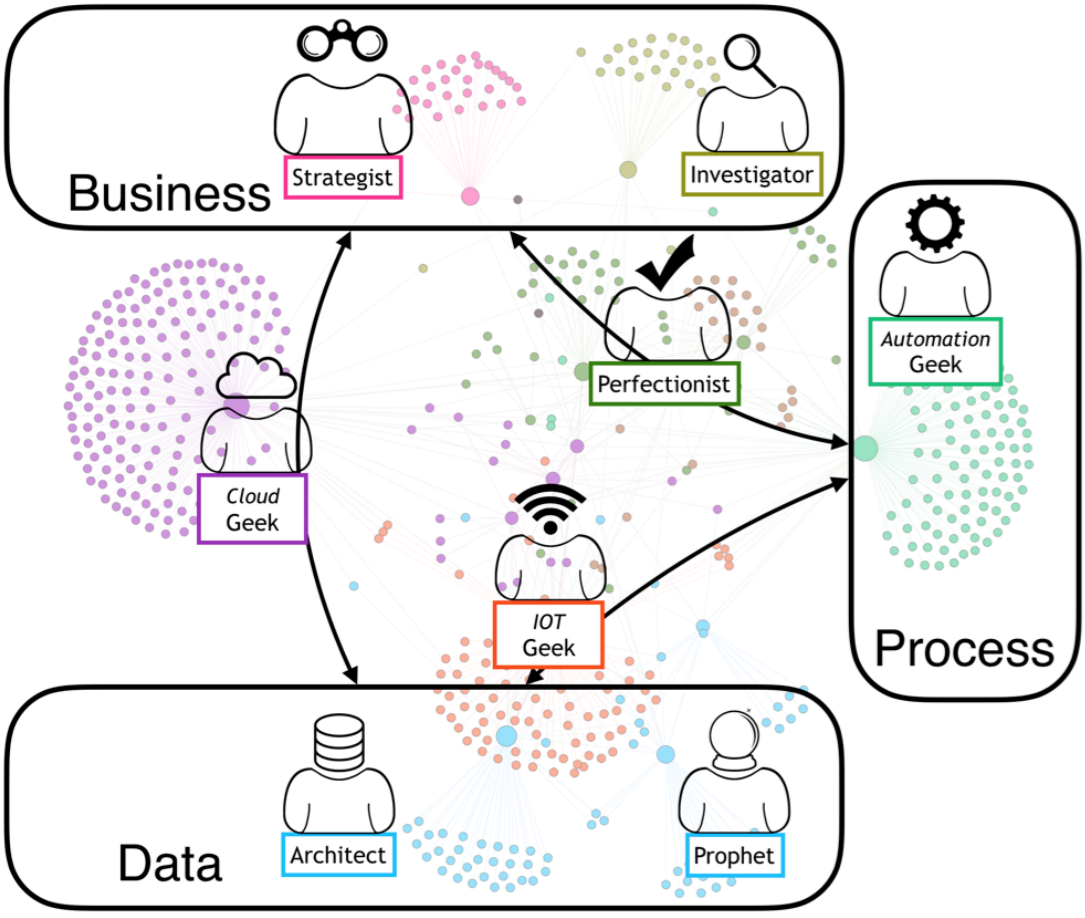

Figure 14.5: The Archetypes grouped by business areas. The image also shows the strategic roles of Perfectionist, Cloud Geek and IOT Geek, who represent the bridges between different functions.

Conclusions and Future Developments

Quali altri stakeholders (quindi testi)

We have finished a nice thesis

Glossary

Generare automaticamente da wikipedia usando tagme.

Morphology= the study of the way words are built up from smaller meaning-bearing units called morphemes.

Bibliography

(1967). Suffix codes for jobs defined in the dictionary of occupational titles, third edition. *United States Employment Service, U.S. Dept. of Labor*.

Abbas, A., Zhang, L., and Khan, S. U. (2014). A literature review on the state-of-the-art in patent analysis. *World Patent Information*, 37:3–13.

Aggarwal, C. C. and Zhai, C. (2012). A survey of text classification algorithms. In *Mining text data*, pages 163–222. Springer.

Ahmed, W. (2017). Using twitter as a data source: an overview of social media research tools (updated for 2017). *Impact of Social Sciences Blog*.

Ahn, H.-I. and Spangler, W. S. (2014). Sales prediction with social media analysis. In *2014 Annual SRII Global Conference (SRII)*, pages 213–222. IEEE.

Alcacer, J. and Gittelman, M. (2006). Patent citations as a measure of knowledge flows: The influence of examiner citations. *The Review of Economics and Statistics*, 88(4):774–779.

Allaire, J., Xie, Y., McPherson, J., Luraschi, J., Ushey, K., Atkins, A., Wickham, H., Cheng, J., and Chang, W. (2018). *rmarkdown: Dynamic Documents for R*. R package version 1.9.

Allen, D. and Wilson, T. D. (2003). Information overload: context and causes. *The New Review of Information Behaviour Research*, 4(1):31–44.

Allwood, J. M. (2014). Squaring the circular economy: The role of recycling within a hierarchy of material management strategies. In *Handbook of recycling*, pages 445–477. Elsevier.

Andersen, B. (1999). The hunt for s-shaped growth paths in technological innovation: a patent study. *Journal of evolutionary economics*, 9(4):487–526.

Anick, P. G. and Vaithyanathan, S. (1997). Exploiting clustering and phrases for context-based information retrieval. In *ACM SIGIR Forum*, volume 31, pages 314–323. ACM.

Apreda, R., Bonaccorsi, A., dell'Orletta, F., and Fantoni, G. (2016). Functional technology foresight. a novel methodology to identify emerging technologies. *European Journal of Futures Research*, 4(1):13.

Arun, R., Suresh, V., Madhavan, C. V., and Murthy, M. N. (2010). On finding the natural number of topics with latent dirichlet allocation: Some observations. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 391–402. Springer.

Asur, S. and Huberman, B. A. (2010). Predicting the future with social media. In *Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology-Volume 01*, pages 492–499. IEEE Computer Society.

Attardi, G. (2006). Experiments with a multilanguage non-projective dependency parser. In *Proceedings of the Tenth Conference on Computational Natural Language Learning*, pages 166–170. Association for Computational Linguistics.

Attardi, G. and Dell'Orletta, F. (2009). Reverse revision and linear tree combination for dependency parsing. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, pages 261–264. Association for Computational Linguistics.

Attardi, G., Dell'Orletta, F., Simi, M., and Turian, J. (2009). Accurate dependency parsing with a stacked multilayer perceptron. *Proceedings of EVALITA*, 9:1–8.

Baccianella, S., Esuli, A., and Sebastiani, F. (2010). Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In *Lrec*, volume 10, pages 2200–2204.

Baker, L. D. and McCallum, A. K. (1998). Distributional clustering of words for text classification. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 96–103. ACM.

Baroncelli, E., Fink, C., and Smarzynska, B. (2004). *The Global Distribution of Trademarks: some stylized facts*. The World Bank.

Bassecoulard, E., Lelu, A., and Zitt, M. (2007). Mapping nanosciences by citation flows: A preliminary analysis. *Scientometrics*, 70(3):859–880.

Bastian, M., Heymann, S., and Jacomy, M. (2009). Gephi: An open source software for exploring and manipulating networks.

Bekkerman, R., El-Yaniv, R., Tishby, N., and Winter, Y. (2001). On feature distributional clustering for text categorization. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 146–153. ACM.

Bellegarda, J. R. (2004). Statistical language model adaptation: review and perspectives. *Speech communication*, 42(1):93–108.

Beller, C., Harman, C., and Van Durme, B. (2014a). Predicting fine-grained social roles with selectional preferences. *ACL 2014*, page 50.

Beller, C., Knowles, R., Harman, C., Bergsma, S., Mitchell, M., and Van Durme, B. (2014b). I'm a belieber: Social roles via self-identification and conceptual attributes. In *ACL (2)*, pages 181–186.

Bergmann, I., Butzke, D., Walter, L., Fuerste, J. P., Moehrle, M. G., and Erdmann, V. A. (2008). Evaluating the risk of patent infringement by means of semantic patent analysis: the case of dna chips. *R&d Management*, 38(5):550–562.

Bikakis, N. (2018). Big data visualization tools. *arXiv preprint arXiv:1801.08336*.

Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., and Hellmann, S. (2009). Dbpedia-a crystallization point for the web of data. *Web Semantics: science, services and agents on the world wide web*, 7(3):154–165.

Blanchard, A. (2007). Understanding and customizing stopword lists for enhanced patent mapping. *World Patent Information*, 29(4):308–316.

Blasi, B., Romagnosi, S., and Bonaccorsi, A. (2018). Do ssh researchers have a third mission (and should they have)? In *The Evaluation of Research in Social Sciences and Humanities*, pages 361–392. Springer.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008.

Bod, R. (2013). *A new history of the humanities: The search for principles and patterns from antiquity to the present*. Oxford University Press.

Bollen, J., Mao, H., and Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of computational science*, 2(1):1–8.

Bonaccorsi, Chiarello, F. and D'amico (2017). Mapping users in patents. towards a new methodology and the definition of a research agenda. In *EPIP 2017 Conference BORDEAUX*.

Bongiovanni, R. and Lowenberg-DeBoer, J. (2004). Precision agriculture and sustainability. *Precision agriculture*, 5(4):359–387.

Bonino, D., Ciaramella, A., and Corno, F. (2010). Review of the state-of-the-art in patent information and forthcoming evolutions in intelligent patent informatics. *World Patent Information*, 32(1):30–38.

Borko, H. and Bernick, M. (1963). Automatic document classification. *Journal of the ACM (JACM)*, 10(2):151–162.

Bornmann, L. (2013). What is societal impact of research and how can it be assessed? a literature survey. *Journal of the American Society for Information Science and Technology*, 64(2):217–233.

Bornmann, L. and Haunschild, R. (2017). Does evaluative scientometrics lose its main focus on scientific quality by the new orientation towards societal impact? *Scientometrics*, 110(2):937–943.

Bornmann, L., Haunschild, R., and Marx, W. (2016). Policy documents as sources for measuring societal impact: how often is climate change research mentioned in policy-related documents? *Scientometrics*, 109(3):1477–1495.

Bornmann, L. and Marx, W. (2014). How to evaluate individual researchers working in the natural and life sciences meaningfully? a proposal of methods based on percentiles of citations. *Scientometrics*, 98(1):487–509.

Boyack, K. W., Small, H., and Klavans, R. (2013). Improving the accuracy of co-citation clustering using full text. *Journal of the American Society for Information Science and Technology*, 64(9):1759–1767.

Bozeman, B. and Sarewitz, D. (2011). Public value mapping and science policy evaluation. *Minerva*, 49(1):1–23.

Brants, T., Popat, A. C., Xu, P., Och, F. J., and Dean, J. (2012). Large language models in machine translation. US Patent 8,332,207.

Brettel, M., Friederichsen, N., Keller, M., and Rosenberg, M. (2014). How virtualization, decentralization and network building change the manufacturing landscape: An industry 4.0 perspective. *International Journal of Mechanical, Industrial Science and Engineering*, 8(1):37–44.

Brown, E. D. (2012). Will twitter make you a better investor? a look at sentiment, user reputation and their effect on the stock market. *Proc. of SAIS*.

Brown, T. and Wyatt, J. (2010). Design thinking for social innovation. *Development Outreach*, 12(1):29–43.

Bryant, S. L., Forte, A., and Bruckman, A. (2005). Becoming wikipedian: transformation of participation in a collaborative online encyclopedia. In *Proceedings of the 2005 international ACM SIGGROUP conference on Supporting group work*, pages 1–10. ACM.

Buchta, C., Kober, M., Feinerer, I., and Hornik, K. (2012). Spherical k-means clustering. *Journal of Statistical Software*, 50(10):1–22.

Bunescu, R. and Paşca, M. (2006). Using encyclopedic knowledge for named entity disambiguation. In *11th conference of the European Chapter of the Association for Computational Linguistics*.

Burgman, M. A. (2015). *Trusting judgements: how to get the best out of experts*. Cambridge University Press.

Burke, P. and Reitzig, M. (2007). Measuring patent assessment quality-analyzing the degree and kind of (in)consistency in patent offices' decision making. *Research Policy*, 36(9):1404–1430.

Butler, F. P. (1969). Rules defining the use of trade terms in patent applications. *J. Pat. Off. Soc'y*, 51:339.

Cakra, Y. E. and Trisedya, B. D. (2015). Stock price prediction using linear regression based on sentiment analysis. In *Advanced Computer Science and Information Systems (ICACSIS), 2015 International Conference on*, pages 147–154. IEEE.

Callon, M., Courtial, J.-P., Turner, W. A., and Bauin, S. (1983). From translations to problematic networks: An introduction to co-word analysis. *Information (International Social Science Council)*, 22(2):191–235.

Cao, J., Xia, T., Li, J., Zhang, Y., and Tang, S. (2009). A density-based method for adaptive lda model selection. *Neurocomputing*, 72(7-9):1775–1781.

Carnevalli, J. A. and Miguel, P. C. (2008). Review, analysis and classification of the literature on qfd—types of research, difficulties and benefits. *International Journal of Production Economics*, 114(2):737–754.

Cascini, G., Fantechi, A., and Spinicci, E. (2004). Natural language processing of patents and technical documentation. In *International Workshop on Document Analysis Systems*, pages 508–520. Springer.

Cascini, G., Russo, D., and Zini, M. (2007). Computer-aided patent analysis: Finding invention peculiarities. In *Trends in computer aided innovation*, pages 167–178. Springer.

Cascini, G. and Zini, M. (2008). Measuring patent similarity by comparing inventions functional trees. In *Computer-aided innovation (CAI)*, pages 31–42. Springer.

Chang, P.-L., Wu, C.-C., and Leu, H.-J. (2009). Using patent analyses to monitor the technological trends in an emerging field of technology: a case of carbon nanotube field emission display. *Scientometrics*, 82(1):5–19.

Chang, W., Cheng, J., Allaire, J., Xie, Y., and McPherson, J. (2017). *shiny: Web Application Framework for R*. R package version 1.0.5.

Chemchem, A. and Drias, H. (2015). From data mining to knowledge mining: Application to intelligent agents. *Expert Systems with Applications*, 42(3):1436–1445.

Chen, C. and Hicks, D. (2004). Tracing knowledge diffusion. *Scientometrics*, 59(2):199–211.

Chen, J., Haber, E. M., Kang, R., Hsieh, G., and Mahmud, J. (2015). Making use of derived personality: The case of social media ad targeting. In *ICWSM*, pages 51–60.

Cheng, X. and Roth, D. (2013). Relational inference for wikification. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1787–1796.

Cheng, Y.-H., Wu, C.-M., Ku, T., and Chen, G.-D. (2013). A predicting model of tv audience rating based on the facebook. In *Social Computing (SocialCom), 2013 International Conference on*, pages 1034–1037. IEEE.

Chiarello, F., Fantoni, G., Bonaccorsi, A., et al. (2017). Product description in terms of advantages and drawbacks: Exploiting patent information in novel ways. In *DS 87-6 Proceedings of the 21st International Conference on Engineering Design (ICED 17) Vol 6: Design Information and Knowledge, Vancouver, Canada, 21-25.08. 2017*, pages 101–110.

Choi, S., Park, H., Kang, D., Lee, J. Y., and Kim, K. (2012). An sao-based text mining approach to building a technology tree for technology planning. *Expert Systems with Applications*, 39(13):11443–11455.

Choi, S., Yoon, J., Kim, K., Lee, J. Y., and Kim, C.-H. (2011). Sao network analysis of patents for technology trends identification: a case study of polymer electrolyte membrane technology in proton exchange membrane fuel cells. *Scientometrics*, 88(3):863.

Chollet, F. (2015). Keras. https://github.com/fchollet/keras.

Cimino, A., Cresci, S., Dell'Orletta, F., and Tesconi, M. (2014). Linguistically-motivated and lexicon features for sentiment analysis of italian tweets. *4th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2014)*, pages 81–86.

Collomb, A., Costea, C., Joyeux, D., Hasan, O., and Brunie, L. (2014). A study and comparison of sentiment analysis methods for reputation evaluation. *Rapport de recherche RR-LIRIS-2014-002*.

Comission, E. (2016). Germany: Industry 4.0. *Digital Transformation Monitor*.

Cooper, D. R. and Allwood, J. M. (2012). Reusing steel and aluminum components at end of product life. *Environmental science & technology*, 46(18):10334–10340.

Council, N. R. et al. (1998). *Visionary manufacturing challenges for 2020*. National Academies Press.

Cowan, R. and Jonard, N. (2004). Network structure and the diffusion of knowledge. *Journal of economic Dynamics and Control*, 28(8):1557–1575.

Cozzens, S. E., Bobb, K., and Bortagaray, I. (2002). Evaluating the distributional consequences of science and technology policies and programs. *Research Evaluation*, 11(2):101–107.

Crain, S. P., Zhou, K., Yang, S.-H., and Zha, H. (2012). Dimensionality reduction and topic modeling: From latent semantic indexing to latent dirichlet allocation and beyond. In *Mining text data*, pages 129–161. Springer.

Crone, S. F. and Koeppel, C. (2014). Predicting exchange rates with sentiment indicators: An empirical evaluation using text mining and multilayer perceptrons. In *Computational Intelligence for Financial Engineering & Economics (CIFEr), 2104 IEEE Conference on*, pages 114–121. IEEE.

Cutting, D. R., Karger, D. R., and Pedersen, J. O. (1993). Constant interaction-time scatter/gather browsing of very large document collections. In *Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 126–134. ACM.

Cutting, D. R., Karger, D. R., Pedersen, J. O., and Tukey, J. W. (2017). Scatter/gather: A cluster-based approach to browsing large document collections. In *ACM SIGIR Forum*, volume 51, pages 148–159. ACM.

Dance, A. (2013). Impact: Pack a punch. *Nature*, 502(7471):397–398.

Davenport, T. H. and Patil, D. (2012). Data scientist. *Harvard business review*, 90(5):70–76.

Davies, R. (2015). Industry 4.0. digitalisation for productivity and growth. *European Parliamentary Research Service*.

Day, G. S., Shocker, A. D., and Srivastava, R. K. (1979). Customer-oriented approaches to identifying product-markets. *The Journal of Marketing*, pages 8–19.

Degryse, C. (2016). Digitalisation of the economy and its impact on labour markets.

Dell'Orletta, F. (2009). Ensemble system for part-of-speech tagging. *Proceedings of EVALITA*, 9:1–8.

Derrick, G., Meijer, I., and Van Wijk, E. (2014). Unwrapping "impact" for evaluation: A co-word analysis of the uk ref2014 policy documents using vosviewer. In *Proceedings of the science and technology indicators conference*, pages 145–154.

Deveaud, R., SanJuan, E., and Bellot, P. (2014). Accurate and effective latent concept modeling for ad hoc information retrieval. *Document numerique*, 17(1):61–84.

Dewulf, S. (2006). Directed variation: Variation of properties for new or improved function product dna, a base for 'connect and develop'. *ETRIA TRIZ Futures*.

Ding, X., Liu, B., and Yu, P. S. (2008). A holistic lexicon-based approach to opinion mining. In *Proceedings of the 2008 international conference on web search and data mining*, pages 231–240. ACM.

Dog, A. (2015). Psychological dictionaty.

Dolan, S. (2008). Six degrees of wikipedia. *Retrieved June*.

Donovan, C. (2011). State of the art in assessing research impact: introduction to a special issue. *Research Evaluation*, 20(3):175–179.

Duflou, J. R., Sutherland, J. W., Dornfeld, D., Herrmann, C., Jeswiet, J., Kara, S., Hauschild, M., and Kellens, K. (2012). Towards energy and resource efficient manufacturing: A processes and systems approach. *CIRP Annals-Manufacturing Technology*, 61(2):587–609.

Eddy, S. R. (1996a). Hidden markov models. *Current opinion in structural biology*, 6(3):361–365.

Eddy, S. R. (1996b). Hidden markov models. *Current opinion in structural biology*, 6(3):361–365.

Ellis, E. (2015). How the usgs uses twitter data to track earthquakes. *Twitter Data Stories, Twitter*.

Engelsman, E. C. and van Raan, A. F. (1994). A patent-based cartography of technology. *Research policy*, 23(1):1–26.

Eppinger, S. D. and Ulrich, K. T. (1995). Product design and development.

Érdi, P., Makovi, K., Somogyvári, Z., Strandburg, K., Tobochnik, J., Volf, P., and Zalányi, L. (2013). Prediction of emerging technologies based on analysis of the us patent citation network. *Scientometrics*, 95(1):225–242.

Ernø-Kjølhede, E. and Hansson, F. (2011). Measuring research performance during a changing relationship between science and society. *Research Evaluation*, 20(2):131–143.

Ernst, H. (2003). Patent information for strategic technology management. *World patent information*, 25(3):233–242.

Evans, D., Bratton, S., and McKee, J. (2010). *Social media marketing: the next generation of business engagement*. John Wiley & Sons.

Eyal, I. and Sirer, E. G. (2018). Majority is not enough: Bitcoin mining is vulnerable. *Communications of the ACM*, 61(7):95–102.

Fantoni, G., Apreda, R., Dell'Orletta, F., and Monge, M. (2013). Automatic extraction of function–behaviour–state information from patents. *Advanced Engineering Informatics*, 27(3):317–334.

Fausto, S., Machado, F. A., Bento, L. F. J., Iamarino, A., Nahas, T. R., and Munger, D. S. (2012). Research blogging: indexing and registering the change in science 2.0. *PloS one*, 7(12):e50109.

Feng, L., Hu, Y., Li, B., Stanley, H. E., Havlin, S., and Braunstein, L. A. (2015). Competing for attention in social media under information overload conditions. *PloS one*, 10(7):e0126090.

Ferragina, P. and Scaiella, U. (2012). Fast and accurate annotation of short texts with wikipedia pages. *IEEE software*, 29(1):70–75.

Few, S. (2012). *Show me the numbers: Designing tables and graphs to enlighten*. Analytics Press.

Frey, C. B. and Osborne, M. A. (2017). The future of employment: how susceptible are jobs to computerisation? *Technological forecasting and social change*, 114:254–280.

Friendly, M. and Denis, D. J. (2001). Milestones in the history of thematic cartography, statistical graphics, and data visualization. *URL http://www. datavis. ca/milestones*, 32:13.

Gabrilovich, E. and Markovitch, S. (2007). Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *IJcAI*, volume 7, pages 1606–1611.

Gantz, J. and Reinsel, D. (2012). The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the far east. *IDC iView: IDC Analyze the future*, 2007(2012):1–16.

Gerken, J. M. and Moehrle, M. G. (2012). A new instrument for technology monitoring: novelty in patents measured by semantic patent analysis. *Scientometrics*, 91(3):645–670.

Gertsis, A. C. and Vasilikiotis, C. (2018). The lisa and socratees© approach for sustainable crop and soil management. In *Eco-friendly Agro-biological Techniques for Enhancing Crop Productivity*, pages 89–110. Springer.

Ghazinoory, S., Ameri, F., and Farnoodi, S. (2013). An application of the text mining approach to select technology centers of excellence. *Technological Forecasting and Social Change*, 80(5):918–931.

Giacomo, O. (2017). Predicting new product success from social network data. *Master Degree Thesis*.

Gibson, K. R., Gibson, K. R., and Ingold, T. (1994). *Tools, language and cognition in human evolution*. Cambridge University Press.

Glänzel, W. and Czerwon, H.-J. (1996). A new methodological approach to bibliographic coupling and its application to the national, regional and institutional level. *Scientometrics*, 37(2):195–221.

Golzio, D. (2012). Wwwwwhow read a patent!

Grangel-González, I., Halilaj, L., Coskun, G., Auer, S., Collarana, D., and Hoffmeister, M. (2016). Towards a semantic administrative shell for industry 4.0 components. In *Semantic Computing (ICSC), 2016 IEEE Tenth International Conference on*, pages 230–237. IEEE.

Griffiths, T. L. and Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National academy of Sciences*, 101(suppl 1):5228–5235.

Griliches, Z. (1981). Market value, r&d, and patents. *Economics letters*, 7(2):183–187.

Guha, S., Rastogi, R., and Shim, K. (1998). Cure: an efficient clustering algorithm for large databases. In *ACM Sigmod Record*, volume 27, pages 73–84. ACM.

Haapala, K. R., Zhao, F., Camelio, J., Sutherland, J. W., Skerlos, S. J., Dornfeld, D. A., Jawahir, I., Clarens, A. F., and Rickli, J. L. (2013). A review of engineering research in sustainable manufacturing. *Journal of Manufacturing Science and Engineering*, 135(4):041013.

Haley, R. I. (1968). Benefit segmentation: A decision-oriented research tool. *The Journal of Marketing*, pages 30–35.

Hall, B. and Helmers, C. (2018). The impact of international patent systems: Evidence from accession to the european patent convention. Technical report, National Bureau of Economic Research.

Hall, B. H., Jaffe, A. B., and Trajtenberg, M. (2001). The nber patent citation data file: Lessons, insights and methodological tools. Technical report, National Bureau of Economic Research.

Hampton, N. (2016). Understanding the blockchain hype: Why much of it is nothing more than snake oil and spin. *Computerworld*, 5.

Han, J., Kamber, M., and Tung, A. K. (2001). Spatial clustering methods in data mining. *Geographic data mining and knowledge discovery*, pages 188–217.

Hand, D. J. (2007). Principles of data mining. *Drug safety*, 30(7):621–622.

Hankamer, J. (1989). Morphological parsing and the lexicon. In *Lexical representation and process*, pages 392–408. MIT Press.

Haupt, R., Kloyer, M., and Lange, M. (2007). Patent indicators for the technology life cycle development. *Research Policy*, 36(3):387–398.

Hauschildt, J. (1991). Towards measuring the success of innovations. In *Technology Management: The New International Language*, pages 605–608. IEEE.

Haustein, S., Bowman, T. D., Holmberg, K., Tsou, A., Sugimoto, C. R., and Larivière, V. (2016). Tweets as impact indicators: Examining the implications of automated "bot" accounts on t witter. *Journal of the Association for Information Science and Technology*, 67(1):232–238.

Haykin, S. and Network, N. (2004). A comprehensive foundation. *Neural Networks*, 2(2004).

Hearst, M. A., Dumais, S. T., Osman, E., Platt, J., and Scholkopf, B. (1998a). Support vector machines. *Intelligent Systems and their Applications, IEEE*, 13(4):18–28.

Hearst, M. A., Dumais, S. T., Osuna, E., Platt, J., and Scholkopf, B. (1998b). Support vector machines. *IEEE Intelligent Systems and their applications*, 13(4):18–28.

Hecklau, F., Galeitzke, M., Flachs, S., and Kohl, H. (2016). Holistic approach for human resource management in industry 4.0. *Procedia CIRP*, 54:1–6.

Hepp, M., Siorpaes, K., and Bachlechner, D. (2007). Harvesting wiki consensus: Using wikipedia entries as vocabulary for knowledge management. *IEEE Internet Computing*, 11(5).

Herbig, P. A. and Kramer, H. (1994). The effect of information overload on the innovation choice process: Innovation overload. *Journal of Consumer Marketing*, 11(2):45–54.

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National academy of Sciences*, 102(46):16569–16572.

Hu, M. and Liu, B. (2004). Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177. ACM.

Iansiti, M. and Lakhani, K. R. (2017). The truth about blockchain. *Harvard Business Review*, 95(1):118–127.

Idris, K. (2008). Wipo intellectual property handbook: Policy, law and use. *Geneva: WIPO publication*, (489).

Ingarao, G. (2017). Manufacturing strategies for efficiency in energy and resources use: The role of metal shaping processes. *Journal of cleaner production*, 142:2872–2886.

ISO, I. (1999). 13407: Human-centred design processes for interactive systems. *Geneva: ISO*.

iView: IDC Analyze the future, I. (2012). bigger digital shadows, and biggest growth in the far east. *IDC Digital Universe Study, EMC*.

Jacobi, D. (1999). La communication scientifique: discours, figures, modèles.

Jaffe, A. B. (2000). The us patent system in transition: policy innovation and the innovation process. *Research policy*, 29(4-5):531–557.

Jaffe, A. B., Trajtenberg, M., and Henderson, R. (1993). Geographic localization of knowledge spillovers as evidenced by patent citations. *the Quarterly journal of Economics*, 108(3):577–598.

Jain, A. K. and Dubes, R. C. (1988). Algorithms for clustering data.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An introduction to statistical learning*, volume 112. Springer.

Jawahir, I. and Bradley, R. (2016). Technological elements of circular economy and the principles of 6r-based closed-loop material flow in sustainable manufacturing. *Procedia Cirp*, 40:103–108.

Jiang, J. J. and Conrath, D. W. (1997). Semantic similarity based on corpus statistics and lexical taxonomy. *arXiv preprint cmp-lg/9709008*.

Jin, G., Jeong, Y., and Yoon, B. (2015a). Technology-driven roadmaps for identifying new product/market opportunities: Use of text mining and quality function deployment. *Advanced Engineering Informatics*, 29(1):126–138.

Jin, X., Wah, B. W., Cheng, X., and Wang, Y. (2015b). Significance and challenges of big data research. *Big Data Research*, 2(2):59–64.

Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. In *European conference on machine learning*, pages 137–142. Springer.

John Walker, S. (2014). Big data: A revolution that will transform how we live, work, and think.

Johnson, P. (2009). 2 hrm in changing organizational contexts. *Strategic HRM*, page 19.

Joung, J. and Kim, K. (2017). Monitoring emerging technologies for technology planning using technical keyword based analysis from patent data. *Technological Forecasting and Social Change*, 114:281–292.

K., U. (2008). Users, experts, and institutions in design. *Handbook of new product development management*, page 421–438.

Kagermann, H. (2006). Industry 4.0 in a global context. *Strategies for Cooperating with International Partners*.

Kalampokis, E., Tambouris, E., and Tarabanis, K. (2013). Understanding the predictive power of social media. *Internet Research*, 23(5):544–559.

Kanninen, S. and Lemola, T. (2006). Methods for evaluating the impact of basic research funding. *Academy of Finland*.

Karame, G. O., Androulaki, E., and Capkun, S. (2012). Double-spending fast payments in bitcoin. In *Proceedings of the 2012 ACM conference on Computer and communications security*, pages 906–917. ACM.

Karki, M. (1997). Patent citation analysis: A policy analysis tool. *World Patent Information*, 19(4):269–272.

Kasemsap, K. (2015). The role of data mining for business intelligence in knowledge management. In *Integration of data mining in business intelligence systems*, pages 12–33. IGI Global.

Kaufman, L. and Rousseeuw, P. J. (2009). *Finding groups in data: an introduction to cluster analysis*, volume 344. John Wiley and Sons.

Khazragui, H. and Hudson, J. (2014). Measuring the benefits of university research: impact and the ref in the uk. *Research Evaluation*, 24(1):51–62.

Kim, G. and Bae, J. (2017). A novel approach to forecast promising technology through patent analysis. *Technological Forecasting and Social Change*, 117:228–237.

Kim, H., Choi, S., Jeong, C., and Kim, K. (2010). Cause-and-effect function analysis. In *Management of innovation and technology (ICMIT), 2010 IEEE international conference on*, pages 518–523. IEEE.

Kim, J. and Lee, S. (2015). Patent databases for innovation studies: A comparative analysis of uspto, epo, jpo and kipo. *Technological Forecasting and Social Change*, 92:332–345.

Kim, S.-M. and Hovy, E. (2004). Determining the sentiment of opinions. In *Proceedings of the 20th international conference on Computational Linguistics*, page 1367. Association for Computational Linguistics.

Kim, Y. (2014). Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*.

Kim, Y., Suh, B., and Lee, K. (2014). # nowplaying the future billboard: mining music listening behaviors of twitter users for hit song prediction. In *Proceedings of the first international workshop on Social media retrieval and analysis*, pages 51–56. ACM.

Kirchmann, H. and Thorvaldsson, G. (2000). Challenging targets for future agriculture. *European Journal of Agronomy*, 12(3-4):145–161.

Kittur, A., Chi, E. H., and Suh, B. (2008). Crowdsourcing user studies with mechanical turk. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 453–456. ACM.

Konkel, F. (2013). Tweets give usgs early warning on earthquakes. *The Business of Federal Technology*.

Kordonis, J., Symeonidis, S., and Arampatzis, A. (2016). Stock price forecasting via sentiment analysis on twitter. In *Proceedings of the 20th Pan-Hellenic Conference on Informatics*, page 36. ACM.

Krallinger, M., Leitner, F., Rabal, O., Vazquez, M., Oyarzabal, J., and Valencia, A. (2015). Chemdner: The drugs and chemical names extraction challenge. *J. Cheminformatics*, 7(S-1):S1.

Kreuchauff, F. and Korzinov, V. (2017). A patent search strategy based on machine learning for the emerging field of service robotics. *Scientometrics*, 111(2):743–772.

Krikorian, R. (2013). New tweets per second record, and how. *Twitter Engineering Blog*, 16.

Kuusi, O. and Meyer, M. (2007). Anticipating technological breakthroughs: Using bibliographic coupling to explore the nanotubes paradigm. *Scientometrics*, 70(3):759–777.

Lafferty, J., McCallum, A., and Pereira, F. C. (2001a). Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

Lafferty, J., McCallum, A., and Pereira, F. C. (2001b). Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

Lafferty, J. and Zhai, C. (2001). Document language models, query models, and risk minimization for information retrieval. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 111–119. ACM.

Lai, K.-K. and Wu, S.-J. (2005). Using the patent co-citation approach to establish a new patent classification system. *Information processing & management*, 41(2):313–330.

Lai, S., Xu, L., Liu, K., and Zhao, J. (2015). Recurrent convolutional neural networks for text classification. In *AAAI*, volume 333, pages 2267–2273.

Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., and Dyer, C. (2016). Neural architectures for named entity recognition. *arXiv preprint arXiv:1603.01360*.

Larose, D. T. and Larose, C. D. (2014). *Discovering knowledge in data: an introduction to data mining*. John Wiley & Sons.

Lasi, H., Fettke, P., Kemper, H.-G., Feld, T., and Hoffmann, M. (2014). Industry 4.0. *Business & Information Systems Engineering*, 6(4):239–242.

Leaman, R., Wei, C.-H., and Lu, Z. (2015). tmchem: a high performance approach for chemical named entity recognition and normalization. *J. Cheminformatics*, 7(S-1):S3.

Lee, C., Kang, B., and Shin, J. (2015a). Novelty-focused patent mapping for technology opportunity analysis. *Technological Forecasting and Social Change*, 90:355–365.

Lee, C., Kang, B., and Shin, J. (2015b). Novelty-focused patent mapping for technology opportunity analysis. *Technological Forecasting and Social Change*, 90:355–365.

Lee, C., Kim, J., Kwon, O., and Woo, H.-G. (2016). Stochastic technology life cycle analysis using multiple patent indicators. *Technological Forecasting and Social Change*, 106:53–64.

Lee, S. and Kim, W. (2017). The knowledge network dynamics in a mobile ecosystem: a patent citation analysis. *Scientometrics*, 111(2):717–742.

Lee, S., Yoon, B., and Park, Y. (2009a). An approach to discovering new technology opportunities: Keyword-based patent map approach. *Technovation*, 29(6–7):481 – 497.

Lee, S., Yoon, B., and Park, Y. (2009b). An approach to discovering new technology opportunities: Keyword-based patent map approach. *Technovation*, 29(6-7):481–497.

Lee, T.-S. and Kang, S.-S. (2014). Optimizing the features of crf-based named entity recognition for patent documents. *Technology*, 18:2–56.

Leek, J. (2015). The elements of data analytic style. *J. Leek.—Amazon Digital Services, Inc.*

León-Rovira, N. and Cho, S. (2007). *Trends in Computer Aided Innovation: Second IFIP Working Conference on Computer Aided Innovation, October 8-9 2007, Michigan, USA*, volume 250. Springer Science & Business Media.

Levitin, D. J. (2014). *The organized mind: Thinking straight in the age of information overload*. Penguin.

Leydesdorff, L. and Hellsten, I. (2006). Measuring the meaning of words in contexts: An automated analysis of controversies about'monarch butterflies,"frankenfoods,'and'stem cells'. *Scientometrics*, 67(2):231–258.

Leydesdroff, L. (1989). Words and co-words as indicators of intellectual organization. *Research policy*, 18(4):209–223.

Li, Y., Rakesh, V., and Reddy, C. K. (2016). Project success prediction in crowdfunding environments. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*, pages 247–256. ACM.

Liaghat, S., Balasundram, S. K., et al. (2010). A review: The role of remote sensing in precision agriculture. *American journal of agricultural and biological sciences*, 5(1):50–55.

Liang, Y. and Tan, R. (2007). A text-mining-based patent analysis in product innovative process. In *Trends in computer aided innovation*, pages 89–96. Springer.

Lih, A. (2004). Wikipedia as participatory journalism: Reliable sources? metrics for evaluating collaborative media as a news resource. *Nature*, 3(1).

Lin, I.-C. and Liao, T.-C. (2017). A survey of blockchain security issues and challenges. *IJ Network Security*, 19(5):653–659.

Liu, B. and Zhang, L. (2012). A survey of opinion mining and sentiment analysis. In *Mining text data*, pages 415–463. Springer.

Liu, H.-C., Liu, L., and Liu, N. (2013). Risk evaluation approaches in failure mode and effects analysis: A literature review. *Expert systems with applications*, 40(2):828–838.

London, K. C. and Science, D. (2015). The nature, scale and beneficiaries of research impact: An initial analysis of research excellence framework (ref) 2014 impact case studies.

Loukides, M. (2011). *What is data science? "* O'Reilly Media, Inc.".

Madani, F. and Weber, C. (2016). The evolution of patent mining: Applying bibliometrics analysis and keyword network analysis. *World Patent Information*, 46:32–48.

Manevitz, L. and Yousef, M. (2007). One-class document classification via neural networks. *Neurocomputing*, 70(7-9):1466–1481.

Manevitz, L. M. and Yousef, M. (2001). One-class svms for document classification. *Journal of machine Learning research*, 2(Dec):139–154.

McTear, M., Callejas, Z., and Griol, D. (2016). *The conversational interface: Talking to smart devices*. Springer.

Medelyan, O., Milne, D., Legg, C., and Witten, I. H. (2009). Mining meaning from wikipedia. *International Journal of Human-Computer Studies*, 67(9):716–754.

MERCER, J., Lopes, T. D. S., and DUGUID, P. (2010). Reading registrations: an overview of 100 years of trademark registrations in france, the united kingdom, and the united states. In *Trademarks, Brands, and Competitiveness*, pages 27–48. Routledge.

Mihalcea, R. and Csomai, A. (2007). Wikify!: linking documents to encyclopedic knowledge. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 233–242. ACM.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Milat, A. J., Bauman, A. E., and Redman, S. (2015). A narrative review of research impact assessment models and methods. *Health Research Policy and Systems*, 13(1):18.

Miller, G. A. (1995). Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Milne, D. (2007). Computing semantic relatedness using wikipedia link structure. In *Proceedings of the new zealand computer science research student conference*.

Milne, D. and Witten, I. H. (2008). Learning to link with wikipedia. In *Proceedings of the 17th ACM conference on Information and knowledge management*, pages 509–518. ACM.

Mining, W. I. D. (2006). Data mining: Concepts and techniques. *Morgan Kaufinann*.

Mirtalaie, M. A., Hussain, O. K., Chang, E., and Hussain, F. K. (2018a). Extracting sentiment knowledge from pros/cons product reviews discovering features along with the polarity strength of their associated opinions. *Expert Systems with Applications*.

Mirtalaie, M. A., Hussain, O. K., Chang, E., and Hussain, F. K. (2018b). Extracting sentiment knowledge from pros/cons product reviews: Discovering features along with the polarity strength of their associated opinions. *Expert Systems with Applications*, 114:267–288.

Misawa, S., Taniguchi, M., Miura, Y., and Ohkuma, T. (2017). Character-based bidirectional lstm-crf with words and characters for japanese named entity recognition. In *Proceedings of the First Workshop on Subword and Character Level Models in NLP*, pages 97–102.

Mitton, C., Adair, C. E., McKenzie, E., Patten, S. B., and Perry, B. W. (2007). Knowledge transfer and exchange: review and synthesis of the literature. *The Milbank Quarterly*, 85(4):729–768.

Moed, H. F. (2006). *Citation analysis in research evaluation*, volume 9. Springer Science & Business Media.

Moehrle, M. (2010). Measures for textual patent similarities: a guided way to select appropriate approaches. *Scientometrics*, 85(1):95–109.

Mogoutov, A. and Kahane, B. (2007). Data search strategy for science and technology emergence: A scalable and evolutionary query for nanotechnology tracking. *Research Policy*, 36(6):893–903.

Montecchi, T., Russo, D., and Liu, Y. (2013). Searching in cooperative patent classification: Comparison between keyword and concept-based search. *Advanced Engineering Informatics*, 27(3):335–345.

Morris, P. S. (2016). Trademarks as sources of market power: Drugs, beers and product differentiation. *JL & Com.*, 35:163.

Morton, S. (2015). Progressing research impact assessment: A 'contributions' approach. *Research Evaluation*, 24(4):405–419.

Mullen, T. and Collier, N. (2004). Sentiment analysis using support vector machines with diverse information sources. In *Proceedings of the 2004 conference on empirical methods in natural language processing*.

Mulrow, E. J. (2002). The visual display of quantitative information.

Narin, F. (1994). Patent bibliometrics. *Scientometrics*, 30(1):147–155.

Noh, H., Jo, Y., and Lee, S. (2015). Keyword selection and processing strategy for applying text mining to patent analysis. *Expert Systems with Applications*, 42(9):4348–4360.

Nutley, S. M., Walter, I., and Davies, H. T. (2007). *Using evidence: How research can inform public services*. Policy press.

of Economic Research, N. B. (2010). Us business cycle expansions and contractions.

of Health, U. D. and Services, H. (2018). Diseases and conditions.

O'Leary, D. E. (2015). Twitter mining for discovery, prediction and causality: Applications and methodologies. *Intelligent Systems in Accounting, Finance and Management*, 22(3):227–247.

O'Neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy*. Broadway Books.

Organization, W. I. P. (2004). *WIPO Intellectual Property Handbook: Policy, Law and Use*. Number 489. World Intellectual Property Organization.

Organization, W. W. I. P. (1971). International patent classification (ipc).

Ozcan, S. and Islam, N. (2017). Patent information retrieval: approaching a method and analysing nanotechnology patent collaborations. *Scientometrics*, 111(2):941–970.

O'Halloran, D. and Kvochko, E. (2015). Industrial internet of things: unleashing the potential of connected products and services. In *World Economic Forum*, page 40.

Pahl, G. and Beitz, W. (2013). *Engineering design: a systematic approach*. Springer Science and Business Media.

Pang, B., Lee, L., et al. (2008). Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval*, 2(1–2):1–135.

Pang, B., Lee, L., and Vaithyanathan, S. (2002). Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 79–86. Association for Computational Linguistics.

Park, H., Kim, K., Choi, S., and Yoon, J. (2013a). A patent intelligence system for strategic technology planning. *Expert Systems with Applications*, 40(7):2373–2390.

Park, H., Ree, J. J., and Kim, K. (2013b). Identification of promising patents for technology transfers using triz evolution trends. *Expert Systems with Applications*, 40(2):736–743.

Park, H., Yoon, J., and Kim, K. (2011a). Identifying patent infringement using sao based semantic technological similarities. *Scientometrics*, 90(2):515–529.

Park, H., Yoon, J., and Kim, K. (2011b). Identifying patent infringement using sao based semantic technological similarities. *Scientometrics*, 90(2):515–529.

Parker, H. (2017). Opinionated analysis development. *PeerJ Preprints*, 5:e3210v1.

Patent, U. S. and amd European Patent Office, T. O. (2014). Cpc annual report 2014.

Pedersen, T. L. (2018). *ggraph: An Implementation of Grammar of Graphics for Graphs and Networks*. R package version 1.0.1.

Petrov, V. (2002). The laws of system evolution. *The TRIZ journal*, 3:9–17.

Philip, J., Dunphy, D. C., Smith, M. S., Ogilvie, D. M., et al. (1966). The general inquirer: a computer approach to content analysis.

Philipp, M. (2006). Patent filing and searching: Is deflation in quality the inevitable consequence of hyper-inflation in quantity? *World Patent Information*, 28(2):117–121.

Pierce, F. J. and Nowak, P. (1999). Aspects of precision agriculture. In *Advances in agronomy*, volume 67, pages 1–85. Elsevier.

Pierpaoli, E., Carli, G., Pignatti, E., and Canavari, M. (2013). Drivers of precision agriculture technologies adoption: a literature review. *Procedia Technology*, 8:61–69.

Pilkington, M. (2016). 11 blockchain technology: principles and applications. *Research handbook on digital transformations*, page 225.

Ponzetto, S. P. and Strube, M. (2007). Knowledge derived from wikipedia for computing semantic relatedness. *Journal of Artificial Intelligence Research*, 30:181–212.

Pop, I. (2006). An approach of the naive bayes classifier for the document classification. *General Mathematics*, 14(4):135–138.

Porter, M. F. (1980). An algorithm for suffix stripping. *Program*, 14(3):130–137.

Prabowo, R. and Thelwall, M. (2009). Sentiment analysis: A combined approach. *Journal of Informetrics*, 3(2):143–157.

Pressman, D. and Stim, R. (2018). *Nolo's Patents for Beginners: Quick & Legal*. Nolo.

Priem, J., Piwowar, H. A., and Hemminger, B. M. (2012). Altmetrics in the wild: Using social media to explore scholarly impact. *arXiv preprint arXiv:1203.4745*.

Provost, F. and Fawcett, T. (2013). Data science and its relationship to big data and data-driven decision making. *Big data*, 1(1):51–59.

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

R Development Core Team (2008). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Rai, A., Patnayakuni, R., and Seth, N. (2006). Firm performance impacts of digitally enabled supply chain integration capabilities. *MIS Quarterly*, 30(2):225–246.

Ramakrishnan, N., Butler, P., Muthiah, S., Self, N., Khandpur, R., Saraf, P., Wang, W., Cadena, J., Vullikanti, A., Korkmaz, G., et al. (2014). 'beating the news' with embers: forecasting civil unrest using open source indicators. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1799–1808. ACM.

Ramshaw, L. A. and Marcus, M. P. (1999). Text chunking using transformation-based learning. In *Natural language processing using very large corpora*, pages 157–176. Springer.

Rao, T. and Srivastava, S. (2012). Analyzing stock market movements using twitter sentiment analysis. In *Proceedings of the 2012 international conference on advances in social networks analysis and mining (ASONAM 2012)*, pages 119–123. IEEE Computer Society.

Reichardt, J. and Bornholdt, S. (2016). Statistical mechanics of community detection. *Strategies for Cooperating with International Partners*.

Resnik, P. (1999). Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *Journal of artificial intelligence research*, 11:95–130.

Reynolds, T. J., Gengler, C. E., and Howard, D. J. (1995). A means-end analysis of brand persuasion through advertising. *International Journal of research in marketing*, 12(3):257–266.

Ricardo, B. and Berthier, R. (2011). Modern information retrieval: the concepts and technology behind search second edition. *Addision Wesley*, 84(2).

Riel, A., Kreiner, C., Macher, G., and Messnarz, R. (2017). Integrated design for tackling safety and security challenges of smart products and digital manufacturing. *CIRP annals*, 66(1):177–180.

Rinker, T. W. (2018). *sentimentr: Calculate Text Polarity Sentiment*. Buffalo, New York. version 2.3.2.

Rip, A. and Courtial, J. (1984). Co-word maps of biotechnology: An example of cognitive scientometrics. *Scientometrics*, 6(6):381–400.

Roblek, V., Meško, M., and Krapež, A. (2016). A complex view of industry 4.0. *Sage Open*, 6(2):2158244016653987.

Rossum, G. (1995). Python reference manual. Technical report, Amsterdam, The Netherlands, The Netherlands.

Rost, K. (2011). The strength of strong ties in the creation of innovation. *Research policy*, 40(4):588–604.

Rotolo, D., Hicks, D., and Martin, B. R. (2015). What is an emerging technology? *Research Policy*, 44(10):1827–1843.

Rüßmann, M., Lorenz, M., Gerbert, P., Waldner, M., Justus, J., Engel, P., and Harnisch, M. (2015). Industry 4.0: The future of productivity and growth in manufacturing industries. *Boston Consulting Group*, 9.

Ruths, D. and Pfeffer, J. (2014). Social media for large studies of behavior. *Science*, 346(6213):1063–1064.

Samuel, G. N. and Derrick, G. E. (2015). Societal impact evaluation: Exploring evaluator perceptions of the characterization of impact under the ref2014. *Research Evaluation*, 24(3):229–241.

Sawanta, M., Urkude, R., and Jawale, S. (2016). Organized data and information for efficacious agriculture using pride™ model.

Schoen, H., Gayo-Avello, D., Takis Metaxas, P., Mustafaraj, E., Strohmaier, M., and Gloor, P. (2013). The power of prediction with social media. *Internet Research*, 23(5):528–543.

Schrijver, R., Poppe, K., and Daheim, C. (2016). Precision agriculture and the future of farming in europe: Scientific foresight study. *Brussels: European Parliament Research Service*.

Schütze, H. and Silverstein, C. (1997). Projections for efficient document clustering. In *ACM SIGIR Forum*, volume 31, pages 74–81. ACM.

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, 34(1):1–47.

Shen, J., Dong, F., and He, W. (2016). Using media-based emotion to predict commodity price. In *Behavioral, Economic and Socio-cultural Computing (BESC), 2016 International Conference on*, pages 1–6. IEEE.

Sievert, C., Parmer, C., Hocking, T., Chamberlain, S., Ram, K., Corvellec, M., and Despouy, P. (2017). *plotly: Create Interactive Web Graphics via 'plotly.js'*. R package version 4.7.1.

Silge, J. and Robinson, D. (2016). tidytext: Text mining and analysis using tidy data principles in r. *The Journal of Open Source Software*, 1(3):37.

Silver, N. (2012). *The signal and the noise: why so many predictions fail–but some don't*. Penguin.

Small, H. (1973). Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for information Science*, 24(4):265–269.

Small, H. (2006). Tracking and predicting growth areas in science. *Scientometrics*, 68(3):595–610.

Small, H. (2013). *The value of the humanities*. Oxford University Press.

Small, H. and Sweeney, E. (1985). Clustering thescience citation index® using co-citations. *Scientometrics*, 7(3-6):391–409.

Smit, J., Kreutzer, S., Moeller, C., and Carlberg, M. (2016). Industry 4.0. *Study for the ITRE Committee, Policy Department A: Economic and Scientific Policy, European Parliament, Brussels*.

Solow, R. M. (1987). We'd better watch out. *New York Times Book Review*, 36.

Stafford, J. V. (2000). Implementing precision agriculture in the 21st century. *Journal of Agricultural Engineering Research*, 76(3):267–275.

Stern, P., Arnold, E., Carlberg, M., Fridholm, T., Rosemberg, C., and Terrell, M. (2013). *Long term industrial impacts of the Swedish competence centres*. VINNOVA.

Strube, M. and Ponzetto, S. P. (2006). Wikirelate! computing semantic relatedness using wikipedia. In *AAAI*, volume 6, pages 1419–1424.

Tang, P., Mollas-Gallart, J., Grimes, S., Barañano, A., and Holmes, P. (2000). Pilot audit of esprit contribution to regional cohesion (cohen). Technical report, mimeo, SPRU, University of Sussex, Falmer, Brighton.

Terragno, P. (1979). Patent as technical literature. *IEEE Trans Prof Commun*, PC-22(2):101–104. cited By 2.

Tey, Y. S. and Brindal, M. (2012). Factors influencing the adoption of precision agricultural technologies: a review for policy implications. *Precision Agriculture*, 13(6):713–730.

Trappey, A. J., Trappey, C. V., Govindarajan, U. H., Sun, J. J., and Chuang, A. C. (2016). A review of technology standards and patent portfolios for enabling cyber-physical systems in advanced manufacturing. *IEEE Access*, 4:7356–7382.

Tseng, Y.-H., Lin, C.-J., and Lin, Y.-I. (2007). Text mining techniques for patent analysis. *Information Processing & Management*, 43(5):1216–1247.

Tuarob, S. and Tucker, C. S. (2013). Fad or here to stay: Predicting product market adoption and longevity using large scale, social media data. In *ASME 2013 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, pages V02BT02A012–V02BT02A012. American Society of Mechanical Engineers.

Tufte, E. R. (2006). *Beautiful evidence*, volume 1. Graphics Press Cheshire, CT.

Tufte, E. R., Goeler, N. H., and Benson, R. (1990). *Envisioning information*, volume 126. Graphics press Cheshire, CT.

Ţurcan, N. (2015). National bibliometric tool–resource for measuring scientific performances. *Information Science*, 11(3):67.

Tuten, T. L. and Solomon, M. R. (2017). *Social media marketing*. Sage.

Ulrich, K. (2003). *Product design and development*. Tata McGraw-Hill Education.

Underwood, S. (2016). Blockchain beyond bitcoin. *Communications of the ACM*, 59(11):15–17.

Van der Aalst, W. M. (2014). Data scientist: The engineer of the future. In *Enterprise interoperability VI*, pages 13–26. Springer.

Van der Meulen, B. and Rip, A. (2000). Evaluation of societal quality of public sector research in the netherlands. *Research Evaluation*, 9(1):11–25.

Van Raan, A. and Tijssen, R. (1993). The neural net of neural network research: An exercise in bibliometric mapping. *Scientometrics*, 26(1):169–192.

Verhaegen, P.-A., D'hondt, J., Vertommen, J., Dewulf, S., and Duflou, J. R. (2009). Relating properties and functions from patents to triz trends. *CIRP Journal of Manufacturing Science and Technology*, 1(3):126–130.

Verspagen, B. (2007). Mapping technological trajectories as patent citation networks: A study on the history of fuel cell research. *Advances in Complex Systems*, 10(01):93–115.

Wang, M.-Y., Chang, D.-S., and Kao, C.-H. (2010). Identifying technology trends for r&d planning using triz and text mining. *R&d Management*, 40(5):491–509.

Wang, S. and Koopman, R. (2017). Clustering articles based on semantic similarity. *Scientometrics*, 111(2):1017–1031.

Wang, Y. and Wang, Y. (2016). Using social media mining technology to assist in price prediction of stock market. In *Big Data Analysis (ICBDA), 2016 IEEE International Conference on*, pages 1–4. IEEE.

WCED, S. W. S. (1987). World commission on environment and development.

Wee, D., Kelly, R., Cattel, J., and Breunig, M. (2015). Industry 4.0-how to navigate digitization of the manufacturing sector. *McKinsey & Company*, 58.

Weller, K. (2015). Accepting the challenges of social media research. *Online Information Review*, 39(3):281–289.

White, H. D. and Griffith, B. C. (1981). Author cocitation: A literature measure of intellectual structure. *Journal of the American Society for information Science*, 32(3):163–171.

Wickham, H. (2016). *ggplot2: elegant graphics for data analysis*. Springer.

Wickham, H., Chang, W., et al. (2008). ggplot2: An implementation of the grammar of graphics. *R package version 0.7, URL: http://CRAN. R-project. org/package= ggplot2*.

Wickham, H. et al. (2014). Tidy data. *Journal of Statistical Software*, 59(10):1–23.

Wickham, H. and Grolemund, G. (2016). *R for data science: import, tidy, transform, visualize, and model data*. " O'Reilly Media, Inc.".

Wickham, H., Hester, J., and Francois, R. (2017). *readr: Read Rectangular Text Data*. R package version 1.1.1.

Wiebe, J. M., Bruce, R. F., and O'Hara, T. P. (1999). Development and use of a gold-standard data set for subjectivity classifications. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 246–253. Association for Computational Linguistics.

Wilkinson, L. (2006). *The grammar of graphics*. Springer Science & Business Media.

Wilson, T., Wiebe, J., and Hoffmann, P. (2005). Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on human language technology and empirical methods in natural language processing*, pages 347–354. Association for Computational Linguistics.

Xu, F. and Keelj, V. (2014). Collective sentiment mining of microblogs in 24-hour stock price movement prediction. In *Business Informatics (CBI), 2014 IEEE 16th Conference on*, volume 2, pages 60–67. IEEE.

Xu, G., Wu, Z., Li, G., and Chen, E. (2015). Improving contextual advertising matching by using wikipedia thesaurus knowledge. *Knowledge and Information Systems*, 43(3):599–631.

Yang, Q., Li, T., and Wang, K. (2004). Building association-rule based sequential classifiers for web-document prediction. *Data mining and knowledge discovery*, 8(3):253–273.

Yli-Huumo, J., Ko, D., Choi, S., Park, S., and Smolander, K. (2016). Where is current research on blockchain technology?—a systematic review. *PloS one*, 11(10):e0163477.

Yoon, B. and Park, Y. (2004). A text-mining-based patent network: Analytical tool for high-technology trend. *The Journal of High Technology Management Research*, 15(1):37–50.

Yoon, J., Choi, S., and Kim, K. (2011). Invention property-function network analysis of patents: a case of silicon-based thin film solar cells. *Scientometrics*, 86(3):687–703.

Yoon, J. and Kim, K. (2011a). An automated method for identifying triz evolution trends from patents. *Expert Systems with Applications*, 38(12):15540–15548.

Yoon, J. and Kim, K. (2011b). Identifying rapidly evolving technological trends for r&d planning using sao-based semantic patent networks. *Scientometrics*, 88(1):213–228.

Yoon, J. and Kim, K. (2012). Detecting signals of new technological opportunities using semantic patent analysis and outlier detection. *Scientometrics*, 90(2):445–461.

Yoon, J., Park, H., and Kim, K. (2013a). Identifying technological competition trends for r&d planning using dynamic patent maps: Sao-based content analysis. *Scientometrics*, 94(1):313–331.

Yoon, J., Park, H., and Kim, K. (2013b). Identifying technological competition trends for r&d planning using dynamic patent maps: Sao-based content analysis. *Scientometrics*, 94(1):313–331.

Youtie, J., Shapira, P., and Porter, A. L. (2008). Nanotechnology publications and citations by leading countries and blocs. *Journal of Nanoparticle Research*, 10(6):981–986.

Yuwono, B. and Lee, D. L. (1996). Search and ranking algorithms for locating resources on the world wide web. In *Data Engineering, 1996. Proceedings of the Twelfth International Conference on*, pages 164–171. IEEE.

Zhang, N., Wang, M., and Wang, N. (2002). Precision agriculture—a worldwide overview. *Computers and electronics in agriculture*, 36(2-3):113–132.

Zhang, T., Ramakrishnan, R., and Livny, M. (1996). Birch: an efficient data clustering method for very large databases. In *ACM Sigmod Record*, volume 25, pages 103–114. ACM.

Zitt, M. and Bassecoulard, E. (2006). Delineating complex scientific fields by an hybrid lexical-citation method: An application to nanosciences. *Information processing & management*, 42(6):1513–1531.