ALMA MATER STUDIORUM · UNIVERSITÀ DI BOLOGNA

CAMPUS DI CESENA
Scuola di Scienze
Corso di Laurea Magistrale in Ingegneria e Scienze Informatiche

A NEW UNSUPERVISED METHODOLOGY OF DESCRIPTIVE TEXT MINING FOR KNOWLEDGE GRAPH LEARNING

Tesi di Laurea in
Text Mining

Relatore
Prof. Gianluca Moro

Co-relatore
Prof.ssa Antonella Carbonaro

Presentata da
Giacomo Frisoni

Terza Sessione di Laurea
Anno Accademico 2018 – 2019

KEYWORDS

Text Mining

Semantic Web

Knowledge Graphs

Rare Diseases

R

While technology is important,
it's what we do with it that truly matters.
- Muhammad Yunus

After going from doctor to doctor,
I tried to think of how the doctors must have felt...
This is what I think: "This woman is presenting this odd disease
that no one knows how to treat, obviously no cure for, and she and
her husband are sitting here looking at me all moon-eyed
desperate for help with this dilemma they have been blind-sided with.
What am I to do? There simply are no set standards for this,
I'm as helpless as she is,
yet she has come to me asking for my help."
- Nutt, 2007

Sommario

Le malattie rare pongono particolari sfide a pazienti, famiglie, caregiver, medici e ricercatori. Attualmente, sono descritte più di 6000 MR (ma se ne stimano fino a 7000) e oltre 350 milioni di persone convivono con esse (5% della popolazione mondiale). A causa della scarsa disponibilità di informazioni e della loro disgregazione, negli ultimi anni si sta assistendo a una forte crescita di comunità di pazienti su piattaforme social quali Facebook. Il lavoro presentato in questa tesi intende estrarre la conoscenza racchiusa nell'elevata quantità di testo non strutturato generato dagli utenti nel tempo, al fine di rappresentarla in una forma organizzata e abilitando un ragionamento logico deduttivo al di sopra di essa. Partendo dalla consapevolezza della necessità di integrare metodologie diverse in domini complessi, la ricerca mostra l'uso combinato di tecniche di Text Mining e Web Semantico, prendendo l'Acalasia Esofagea come caso di studio. In particolare, si propone la creazione di un'ontologia atta a estendere ORDO e a introdurre una visione incentrata sul paziente nel mondo dei Linked Data. L'importanza di questo studio è da ricercarsi nel fatto che esso costituisca potenzialmente la base di un progetto capace di consentire un rapido accesso a molte informazioni di alto valore (in categorie quali epidemiologia, sintomatologia, diagnosi, trattamenti, farmaci, nutrizione, stile di vita), rispondendo agli interrogativi dei pazienti e fornendo loro un ulteriore strumento attraverso il quale prendere decisioni. Il tutto minimizzando i costi grazie a un reperimento automatico di tali dati e aumentando la produttività dei ricercatori.

Abstract

Rare diseases pose particular challenges to patients, families, caregivers, clinicians and researchers. Currently, more than 6000 rare diseases are described (but up to 7000 are estimated) and more than 350 million people live with them (5% of the world population). Due to the scarce availability of information and their disintegration, in recent years we are witnessing strong growth of patient communities on social platforms such as Facebook. The work presented in this thesis is intended to extract knowledge from the large availability of unstructured text generated by the users over time, in order to represent it in an organized way and to make logical reasoning above. Starting from the awareness of the need to integrate different methodologies in complex domains, the research shows a combined use of Text Mining and Semantic Web techniques, taking Esophageal Achalasia as a case study. In particular, an ontology is created to extend ORDO and introduce a patient-centered vision into the world of linked data. The significance of this development is that it potentially constitutes the basis of a project that can allow rapid access to many high-value information (in topics such as symptomatology, epidemiology, diagnosis, treatments, drugs, nutrition, lifestyle), responding to patients' questions and providing them with an additional tool for decision making, minimizing costs through the automatic retrieval of these data and increasing the productivity of investigators.

Ringraziamenti

Al termine di questo percorso di studi ritengo opportuno e doveroso dedicare qualche riga di ringraziamento a tutti coloro che mi hanno accompagnato lungo il tragitto di questo viaggio.

In primis desidero ringraziare il mio relatore Gianluca Moro per l'attenzione, la disponibilità e la competenza con cui mi ha guidato passo dopo passo nella realizzazione di questo lavoro. La sua esperienza e la sua professionalità si sono rivelate fondamentali. Similmente ringrazio la mia co-relatrice Antonella Carbonaro per la gentilezza, la cortesia, l'intuito e la capacità di condurmi a riflettere sul proseguo del mio percorso verso il mondo della ricerca.

Un vivo ringraziamento al professor Dario Maio, primo docente ad aver appoggiato il mio lavoro fin dallo stadio embrionale e pilastro dei miei progetti universitari.

Grazie a mia madre e a mio padre per il loro amore, per i loro insegnamenti, per avere temprato il mio carattere e per aver infuso in me la consapevolezza che risultati ottimali si ottengono solo con impegno e passione.

Ringrazio la mia intera famiglia per l'affetto incondizionato con cui mi accoglie ogni giorno, anche quando tutto appare complicato e irraggiungibile.

Un grazie di cuore ad Althea per amarmi così come sono con i miei pregi e i miei difetti, preziosa ascoltatrice di sfoghi e turbamenti giornalieri e amatissima confidente.

Non posso dimenticare i miei amici per la complicità dei vissuti, le risate, le angosce, le pacche sulle spalle. Siete parte integrante di me e delle mie mattine, delle corse per il treno, dei ritardi, dei fogli stropicciati, delle nuvole di fumo dalle bocche in inverno, dei fine settimana spensierati. Vi voglio un mondo di bene.

Ringrazio vivamente Marcin per essere un vero amico, il migliore, che ascolta e non giudica, che è sempre presente anche quando non c'è, che apprezzo totalmente per la sua lealtà e semplicità.

Grazie ai miei colleghi di Amae, in particolar modo Celeste, per avermi indotto a comprendere il valore della solidarietà e del sociale e quanto arricchisca il cuore mettersi a disposizione degli altri.

Mille grazie a Birillo, al suo sguardo dolce, al suo tartufo umido appoggiato sulle mie gambe durante le ore di studio, alle sue zampate indagatrici. Sei il mio antistress e amico peloso preferito.

Ringrazio anche tutti coloro che hanno camminato con me anche per poco tempo e che si sono allontanati quando iniziava la salita. Se sono la persona di oggi lo devo anche a voi.

E soprattutto grazie a me, alla mia cocciutaggine, alla mia perseveranza, alle notti di studio, al mio crederci sempre e comunque, alla mia rarità... Questa è solo l'introduzione, ora arriva il Capitolo I.

Giacomo
11 Marzo 2020

Acknowledgments

At the end of this course of study I think it's appropriate and right to dedicate a few lines of thanks to all those who have accompanied me along this adventure.

First of all, I'd like to express my deepest gratitude to my supervisor Gianluca Moro, who guided me step by step in the realization of this work. I'll always be grateful for his care, availability and competence. His experience and professionalism turn out to be fundamental. At the same time I'd like to thank my co-supervisor Antonella Carbonaro for her kindness, courtesy and insight. She led me to consider the continuation of my journey into the world of research.

A sincere thanks to professor Dario Maio, who has been the first one to support my work since the embryonic stage. He's been the pillar of my university projects.

Thanks to my mother and my father for their love, for their teachings and for tempering my character. They are the ones who have always instilled in me the awareness that optimal results can be achieved only through commitment and passion.

I'd like to thank my whole family for the unconditional love they show me every day, even when everything seems complicated and unreachable.

An heartfelt thanks to Althea for loving me exactly as I am, with my strengths and weaknesses. She's the precious listener of my daily outbursts and apprehensions, moreover she's my beloved confidant.

I can't forget to thank my friends for their complicity, the laughter, for the anguish we've always shared and for their unfailing pats on the back. You are an integral part of me and my mornings. We've shared a lot: the rides to catch the train, the delays, the crumpled sheets, the clouds of smoke from the mouths in winter due to the cold, the carefree weekends. I love you so much.

Many thanks to Marcin for being a true friend, the best one. He listens to me without judging, he's always present even when he is not there, I totally appreciate him for his loyalty and simplicity.

Thanks to my colleagues from Amae, in particular to Celeste, for making me understand the importance of solidarity and social field. They show me

how enriching it is to make yourself available to others.

Many thanks to Birillo, to his sweet look, to his wet nose resting on my legs during my study hours, to his inquiring paws. You are my stress-relieving and my best furry friend.

I'd also like to thank all those who walked with me even for a short period of time and who walked away when the climb started. If I am the person you see today I owe it to you too.

And above all thanks to me, to my stubbornness, my perseverance, to the sleepless nights spent studying, to my believing it always and however, to my rarity... This is just the introduction, the following is the first chapter.

Giacomo
March 11, 2020

Introduzione

Motivazioni

Quando si riflette sul concetto di rarità in ambito sanitario, si tende a pensare a un'eventualità remota e riservata a pochi individui. Nel momento in cui questo fenomeno irrompe nella quotidianità di una persona, si realizza come la parola "rara" sia in realtà usata frequentemente in maniera impropria.

Una malattia si definisce rara quando la sua prevalenza, intesa come il numero di casi presenti su una data popolazione, non supera una soglia stabilita (codificata dalla legislazione di ogni singolo paese).

Le malattie rare (abbreviate *MR*) pongono pertanto particolari sfide a pazienti, famiglie, caregiver, medici e ricercatori. Dopo essere state trascurate per molti anni (arrivando ad acquisire la denominazione "*health orphans*"), oggi costituiscono un importante problema di salute pubblica.

Attualmente, più di 6000 malattie rare sono registrate su *Orphanet* [1] (ma se ne stimano fino a 7000 [2]) e oltre 350 milioni di pazienti convivono con esse (5% della popolazione mondiale) [2]. Di conseguenza, se la percentuale di vite coinvolte da una singola MR è da considerarsi per definizione limitata, non lo è il numero di disturbi rientranti in tale categoria: le indicazioni riportate all'interno delle pubblicazioni scientifiche in merito alla numerosità di queste patologie cresce regolarmente con l'avanzare della scienza medica e della ricerca. L'assenza di dati epidemiologici per la maggior parte dei disturbi rende tuttavia difficile la stima del loro reale impatto [3].

Inoltre, la quasi totalità delle malattie rare non ha cura e soltanto il 5% di esse ha opzioni terapeutiche. L'80% è di origine genetica e il 50% delle persone affette da MR si riferisce a bambini in età pediatrica (con un tasso di mortalità nei primi cinque anni di vita pari al 35%) [4, 2, 5]. In aggiunta a questo, tali patologie sono spesso non diagnosticate o diagnosticate erroneamente per periodi prolungati, causando lunghi ritardi diagnostici (con una media globale di 4,8 anni e 7,3 medici consultati, per singola MR) capaci di aumentare significativamente il carico della malattia stessa [6, 7].

A causa della scarsa disponibilità di informazioni e della loro disgregazione, negli ultimi anni si sta assistendo a una forte crescita di comunità di pazienti sul web (in correlazione con l'incoraggiante risultato di una recente indagine compiuta da EURORDIS, che vede i pazienti affetti da malattie rare più propensi a condividere i propri dati sanitari rispetto alla popolazione generale [8]). Il bisogno di dialogare con altre persone aventi la medesima problematica è spesso una naturale conseguenza della volontà di abbattere le barriere di isolamento in cui si è rinchiusi. I contesti social (quali i gruppi *Facebook*) divengono pertanto il territorio attraverso cui si condividono esperienze, si richiedono pareri e si scambiano informazioni di indubbia rilevanza durante tutto il percorso di un malato raro (dalla sintomatologia alla diagnosi, dai trattamenti terapeutici ai centri specializzati, dai medici di riferimento all'impatto sullo stile di vita).

Contributi

La condivisione dei dati sanitari per far avanzare la ricerca scientifica e migliorare i benefici clinici è di particolare importanza nel campo delle malattie rare, in cui le conoscenze e le competenze sono limitate e le popolazioni di pazienti sono geograficamente disperse. Questa tesi nasce dalla comprensione del valore di tali dati accumulati nel tempo e dalla volontà di renderli facilmente disponibili. Nello specifico, si intende estrarre - con un approccio non supervisionato - la conoscenza racchiusa nell'elevata quantità di testo non strutturato generato dagli utenti attraverso interazioni social, al fine di rappresentarla in una forma organizzata e abilitando un ragionamento logico deduttivo al di sopra di essa.

Partendo dalla consapevolezza della necessità di integrare metodologie diverse in domini complessi, la ricerca mostra l'uso combinato di tecniche di *Text Mining* e *Web Semantico*, prendendo l'*Acalasia Esofagea* (*ORPHA:930*) come caso di studio.

La tesi propone una innovativa metodologia di text mining descrittivo, oltre che la creazione di un knowledge graph atto a estendere la *Orphanet Rare Disease Ontology (ORDO)* e a introdurre una visione incentrata sul paziente nel mondo dei *Linked Data*. Combinando un approccio simbolico (per la concettualizzazione della conoscenza degli utenti) con tecniche quantitative all'avanguardia e di carattere originale, la tesi propone come ulteriore contributo anche un nuovo approccio di *Knowledge Graph Learning*.

Impatto sociale

Il lavoro descritto in questo documento costituisce pertanto la base di un progetto capace di consentire un rapido accesso a molte informazioni di alto valore (in categorie quali epidemiologia, sintomatologia, diagnosi, trattamenti, farmaci, nutrizione), rispondendo agli interrogativi dei pazienti e fornendo loro un ulteriore strumento attraverso il quale operare decisioni che richiedano approfondimenti e molteplici punti di vista. Il tutto minimizzando i costi grazie a un reperimento automatico dei dati e aumentando la produttività dei ricercatori (che potrebbero sfruttare tale strumento per aver accesso istantaneo a informazioni provenienti dai pareri diretti di una vasta popolazione di malati e caregiver). La potenzialità dei dati raccolti dai singoli pazienti, aggregati in forma anonima e discussi su così larga scala, creano infatti un'unica opportunità di unire il contesto reale al mondo della ricerca scientifica, contribuendo alla rivoluzione della *digital health*.

Organizzazione della tesi

Il lavoro di tesi è organizzato nei seguenti capitoli:

- **Capitolo 1** - Presenta un quadro generale sui concetti di comprensione e di conoscenza, giungendo a definire come questi vengano applicati sul contenuto testuale. Inoltre, illustra la recente tendenza ad adottare soluzioni di Intelligenza Artificiale che integrino approcci qualitativi e quantitativi.

- **Capitolo 2** - Discute brevemente il background teorico su cui si basa il contributo di Text Mining, focalizzandosi in particolare sulla Latent Semantic Analysis.

- **Capitolo 3** - Affronta lo stato dell'arte legato alle tecniche di Web Semantico trattate all'interno della tesi, approfondendo in particolare i temi dei knowledge graph, delle ontologie probabilistiche e dei progetti attualmente esistenti in ambito malattie rare.

- **Capitolo 4** - Illustra il contributo di Text Mining. Nello specifico, analizza le soluzioni già esistenti in letteratura ponendo particolare attenzione sul concetto di explainability. Successivamente presenta un nuovo metodo non supervisionato di analisi descrittiva.

- **Capitolo 5** - Propone il contributo di Web Semantico. Partendo dalla conoscenza estratta attraverso il metodo di text mining descrittivo, affronta la problematica della sua rappresentazione automatica per mezzo di una nuova metodologia di Knowledge Graph Learning.

- **Capitolo 6** - Entrambi i contributi vengono applicati a un caso di studio in ambito medico focalizzato sull'Acalasia Esofagea. Si descrivono aspetti implementativi e si valutano i risultati ottenuti.

Introduction

Motivation

When reflecting on the concept of rarity in the health sector, we tend to think of a remote event reserved for a few individuals. The moment this phenomenon breaks into a person's daily life, it is realized how the word "rare" is actually frequently used improperly.

A disease is defined as rare when its prevalence, understood as the number of cases present in a given population, does not exceed a set threshold (codified by the legislation of each individual country).

Rare diseases (shortened as *RDs*) therefore pose particular challenges to patients, families, caregivers, clinicians and researchers. After being neglected for many years (coming to be called "*health orphans*"), today they are an important public-health issue.

Currently, more than 6000 rare diseases are registered on *Orphanet* [1] (but up to 7000 are estimated [2]) and more than 350 million people live with them (5% of the world population) [2]. Consequently, if the percentage of lives involved by a single RD is to be considered limited by definition, the number of disorders that fit this category is not: the indications reported in the scientific publications regarding the number of these diseases grow regularly with the advancement of medical science and research. The absence of epidemiological data for most RDs, however, makes it difficult to estimate their real impact [3].

Moreover, there is no cure for the vast majority of rare diseases and only 5% of them have treatment options. 80% of RDs are of genetic origin and 50% affect children (with a mortality rate in the first five years of life equal to 30%) [4, 2, 5]. In addition to this, these disorders are often un- or misdiagnosed for extended periods, resulting in a long diagnostic delay (with a global average of 4.8 years and 7.3 consulted physicians for RD) that may significantly add to the burden of the disease itself [6, 7].

Due to the scarce availability of information and their disintegration, in recent years we are witnessing a strong growth of patient communities on the web (in correlation with the encouraging result of a recent survey carried out by

EURORDIS, which sees rare disease patients more likely to share their health data than the general population [8]). The need to communicate with other people with the same problem is often a natural consequence of the desire to break down the isolation barriers in which a patient is locked up. Social contexts (such as *Facebook* groups) therefore become the environment through which experiences are shared, opinions are requested and information of undoubted relevance is exchanged throughout the whole rare patient path (from symptoms to diagnosis, from therapeutic treatments to specialized centers, from reference doctors to lifestyle impact).

Contribution

Sharing health data to advance scientific research and improve clinical benefits are of particular importance in the field of rare diseases where knowledge and expertise are limited and patient populations are geographically dispersed. This thesis arises from an understanding of the value of such data accumulated over time and from the desire to make them easily available. Specifically, it is intended to extract - with a semi-supervised approach - the knowledge contained in the high amount of unstructured text generated by users through social interactions, in order to represent it in an organized way and enabling deductive logical reasoning above it.

Starting from the awareness of the need to integrate different methodologies in complex domains, the research shows the combined use of *Text Mining* and *Semantic Web* techniques, taking the *Esophageal Achalasia* (*ORPHA: 930*) as a case study.

The thesis proposes an innovative descriptive text mining method, as well as the creation of a knowledge graph aimed at extending the *Orphanet Rare Disease Ontology (ORDO)* and to introduce a patient-centered vision into the world of *Linked Data*. Combining a symbolic approach (for the conceptualization of user knowledge) with cutting-edge quantitative techniques of an original character, the thesis proposes as a further contribution also a new approach of *Knowledge Graph Learning*.

Social Impact

The work described in this document forms the basis of a project capable of allowing rapid access to many high-value information (in topics such as epidemiology, symptomatology, diagnosis, treatments, drugs, nutrition), responding to patients' questions and providing them with an additional tool for decision making where insights and multiple points of view are required. All this by minimizing costs thanks to automatic data retrieval and increasing the productivity of researchers (who could take advantage of this tool to have instant access to knowledge coming from the direct opinions of a large population of patients and caregivers). The potential of the data collected by individual patients, aggregated anonymously and discussed on such large scale, create a unique opportunity to combine the real context with the world of scientific research, contributing to the *digital health* revolution.

Thesis Organization

The thesis is organized as follows.

- **Chapter 1** - Presents a general framework on the concepts of understanding and knowledge, coming to define how these are applied to textual content. Furthermore, it illustrates the recent trend to adopt Artificial Intelligence solutions that integrate qualitative and quantitative approaches.

- **Chapter 2** - Briefly discusses the theoretical background on which the contribution of Text Mining is based, focusing in particular on the Latent Semantic Analysis.

- **Chapter 3** - Describes the state-of-the-art linked to the Semantic Web techniques dealt inside the thesis, focusing in particular on knowledge graphs, probabilistic ontologies and projects currently existing in the field of rare diseases.

- **Chapter 4** - Illustrates the contribution of Text Mining. Specifically, it analyzes the solutions already existing in the literature paying particular attention to the concept of explainability. He then presents a new unsupervised method of descriptive analysis.

- **Chapter 5** - It proposes the contribution of the Semantic Web. Starting from the knowledge extracted through the descriptive text mining method, it tackles the problem of its automatic representation by means of a new method of Knowledge Graph Learning.

- **Chapter 6** - Both contributions are applied to a medical case study focused on Esophageal Achalasia. Implementation aspects are described and the results obtained are evaluated.

Contents

List of Figures

List of Tables

Chapter 1

Understanding and Knowledge

> *"And even in our wildest and most wandering reveries, nay in our very dreams,*
> *we shall find, if we reflect, that the imagination ran not altogether at adventures,*
> *but that there was still a connection upheld among the different ideas,*
> *which succeeded each other. Were the loosest and freest conversation to be transcribed,*
> *there would immediately be transcribed, there would immediately be observed*
> *something which connected it in all its transitions."*
> *- David Hume, An enquiry concerning human understanding, 1748*

Knowledge understanding and knowledge representation are the major topics covered by this dissertation. To be able to explore the possibilities related to them, it is important to understand the foundation of the underlying concepts. This chapter presents an overview of the various topics behind the importance of meaning, motivating some of the choices made in the thesis.

1.1 Basic concepts

This section illustrates the fundamental elements to understand the concept of understanding, on which both text mining and the semantic web are built. The logical flow adopted for the illustration of the concepts is inspired by the "Information Service Engineering" course of Dr. Harald Sack and Dr. Maria Koutraki.

1.1.1 Communication

According to Merriam-Webster dictionary, *"communication is a process by which information is exchanged between individuals through a common system of* *symbols, signs, or behavior"*. In a similar way, Wikipedia defines communication as *"the act of conveying intended meanings from one entity or group to another through the use of mutually understood signs and semiotic rules"*. Therefore symbols, signs and behaviors characterize a language, while semiotic rules tell us how to interpret language's items.

In a communication process usually there is some information that has to be encoded in terms of language and passed from one sender to a receiver, who has to decode the message from the language to obtain the meaning originally intended by the sender. So, at the beginning of the communication the sender forms a verbal utterance from a thought, following specific rules of shared syntax and semantics to ensure that the recipient will be able to reconstruct the content of the thought itself [9]. Figure 1.1 shows a schematic representation of the communication process, based on the information theory.

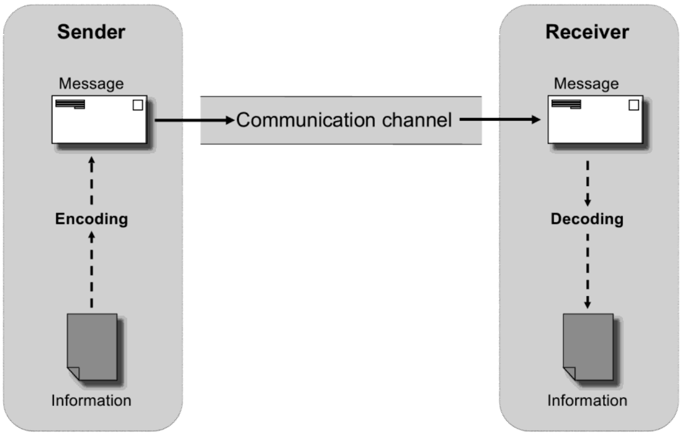


Figure 1.1: Communication model seen from the information theory perspective
From "Digital Communication: Communication, Multimedia, Security", 2014 [9]

1.1.2 Language

Language is defined by Encyclopaedia Britannica as *"a system of conventional spoken, manual, or written symbols by means of which human beings, as members of a social group and participants in its culture, express themselves. The functions of language include communication, the expression of identity, play, imaginative expression, and emotional release"*.

In Natural Language Processing (NLP) we are obviously interested in natural languages. A natural language (or ordinary language) is any language that has evolved naturally in humans through use and repetition without conscious planning or premeditation (unlike formal languages such as programming ones). Textual messages belong to this category.

1.1.3 Information

Many definitions of information have been proposed and each of these focuses on a specific aspect, giving its own value to the total picture.
Overall, it can be said that information is that which informs, and is conveyed as the content of a message that is expressed with the help of language.
Information can be encoded into various forms for transmission and interpretation. In a physical sense, information is any propagation of cause and effect within a system (so it does not necessarily depend on natural language).
Another aspect is that information's existence is not necessarily coupled to an observer. However, from a statistical point of view, the most important definition of information is that *"information reduces uncertainty"*.

1.1.4 Understanding

Understanding is the ability to grasp the meaning of information. So, to understand the content of a message the receiver must be able to interpret it correctly. In other words, understanding is always the correct interpretation of the meaning of a message.

Correct interpretation depends on the following involved factors.

- **Syntax**
  The most basic building block of information of languages.
  Syntax is a word which stems from the Greek for arrangement or ordering.
  In grammar syntax deals with the relations between words in sentence structure.
  In formal languages (not natural languages), syntax is just a set of rules, by which well-formed expressions can be created from a fundamental set of symbols (alphabet). So rules determine how the symbols of an alphabet have to be put together to form something meaningful.
  Syntax only tells what kind of arrangement of characters are allowed and which not.

- **Semantics**
  Semantics is a term which also come from Greek and it represents the study of meaning. Semantics defines how the symbols created in the syntax will be interpreted. So it is focused on sense and meaning of language or symbols provided by language.
  It is a fundamental concept for understanding. In summary, it is the study of the interpretation of signs or symbols as used by agents or communities within particular circumstances and contexts.
  The semantics of a message depends on context and pragmatics.

It is interesting to observe that semantics works with syntax, deriving sometimes the meaning of complex concepts from the meaning of more basic concepts precisely through syntactic rules.

- **Context**
  Context come from Latin and means “interweaved”.
  It denotes the surrounding of a symbol (concept) in an expression respect its relationship with surrounding expressions (concepts) and further related elements.
  Context is important to disambiguate or to define the right semantics or the right way to interpret a message. From a certain point of view, it defines a time frame in which a specific semantics is valid. So, the context is constituted by all the additional information that comes with the message, defining which kind of semantics will be chosen.
  A distinction is made between general contexts (i.e., place, time) and personal or social contexts (i.e., relation between sender and receiver of a message).

- **Pragmatics**
  In Greek, pragmatics means action. Pragmatics reflects the intention of the sender of a message, and so the language application in different situations.

- **Experience**
  A message is formulated based on the experience of the sender and the recipient regarding the world knowledge. Indeed, one’s knowledge of the world also determines how a message is interpreted.

In conclusion, there are lots of factors which influence the correct interpretation. This is also one of the reasons why a computer’s understanding of natural language is not a simple task. Natural languages are highly ambiguous and usually take into account different possible contexts. Understanding formal languages is much easier than natural ones simply because of the lack of ambiguity.

1.1.5 Knowledge

Knowledge is the very focus of this thesis and its nature needs to be analysed. Knowledge is a very powerful abstract concept without any reference to the tangible world. However it does not yet have a clear definition. Identifying a way to represent knowledge has been an important challenge since the dawn of the human race. The creation of written and spoken languages is the foremost example of the effort to represent knowledge in such a way as to preserve it

and to guarantee that it can be transmitted to future generations.
Starting with the Greek philosophers, many attempts to define knowledge have followed over time. Even today, due to the many aspects related to the knowledge itself, finding a universally accepted definition is not easy. Consequently, there are many versions of it.
In a broad sense, knowledge is information possessed in the mind of an individual (personalized or subjective). It can be related to facts, procedures, concepts, ideas, observations, interpretations and judgments (which may or not be unique, useful, accurate, or structured).
Knowledge is defined by the lexicographers at Cambridge Dictionary as *"understanding of or information about a subject that you get by experience or study, either known by one person or by people generally"*. Vice versa, the classic and most widespread definition is that provided by the philosopher Plato, who sees knowledge as *"justified true belief"*.
A good analysis of the problem is reported in [10].
The terms "data", "information", "knowledge" and "wisdom" are often confused and overlapping, despite having different meanings. Clarifying this terminological distinction is an effective way to better understand the concept of knowledge itself. The Data-Information-Knowledge-Wisdom (DIKW) hierarchy or pyramid, depicted in Figure 1.2, is a widely used model within information science and knowledge management. The main interpretation of the DIKW

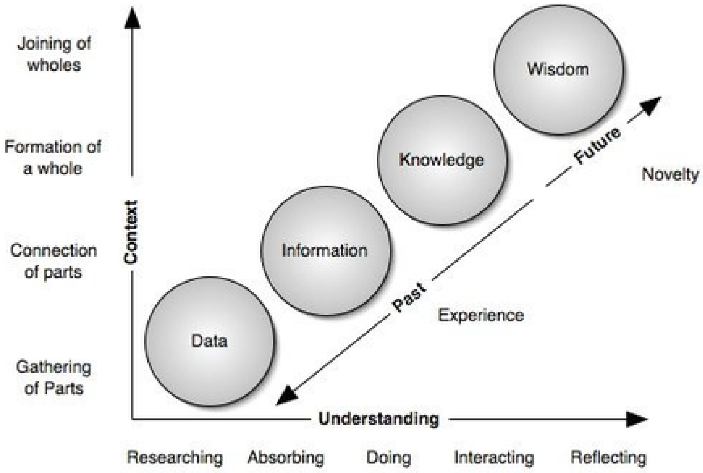


Figure 1.2: A detailed illustration of the DIKW hierarchy
From "The Application of Visual Analytics to Financial Stability Monitoring", 2014

hierarhcy is that large amounts of data (pure and simple raw facts obtained by observation and without a particular organization) are distilled to a smaller quantity of information (structured, linked and meaningful data). Then, a still rather large amount of information is further processed to create a more limited knowledge (connected, organized, contextualized, understood and query-ready information). Finally, wisdom (distilled knowledge and understanding which can lead to decisions) is built on the top of the pyramid.

1.2 Understanding Textual Content

There is a precondition for enabling machines to understand natural language: the textual content must be read and interpreted correctly by them. To do this there are mainly two distinct approaches.

- **Quantitative (Statistical AI)**
  Methods that try to understand or to learn the content, using statistical models, machine learning, deep learning and other natural language processing techniques, conveying the semantics in an implicit way. The language is thus learned from its usage.
- **Qualitative (Symbolic AI)**
  Methods that add explicit semantic information to documents.
  The Semantic Web is the main example of this category. Specifically, the Semantic Web annotate the natural language web content with semantic metadata that encode the meaning of the content itself and allow the correct interpretation by machines.
  In this context, a simple class hierarchy with only "is_a" relations is not sufficient for a complete understanding. For this reason, properties (having a domain and a range) are also required in order to model how concepts bind to other knowledge.
  Implicit hidden knowledge can be deduced from the explicitly available one with the help of logical inference.
  The ontologies are standardized knowledge representations which explicit the meaning of information (Semantics) in a formal and structured way, making possible to:
  - process it automatically;
  - integrate heterogeneous data;
  - deduce not evident new information.

1.2.1 Towards a mixed AI

Today most of the works fall into one of these two categories.
The consideration of mixed solutions has started to develop only recently.
One of the objectives of the research described in this document is precisely that

of combining quantitative and qualitative techniques. In doing this, we want to be careful to merge the advantages of both approaches, therefore making wise use of the methodologies that characterize these two branches of Artificial Intelligence (AI).

From the articles published in the literature, it is easy to deduce how the union of Statistical AI and Symbolic AI constitutes a winning approach and how one strengthens the other. In fact, research is currently being done in both directions.

- **Statistical AI → Symbolic AI**
  Statistical AI techniques can be used to extract knowledge from unstructured text, facilitating the construction and population of ontologies capable of representing it (enabling Symbolic AI). This is the one on which the thesis is focused.

- **Statistical AI ← Symbolic AI**
  Integrated and linked data can be used as a more expressive semantic data model to feed the ML algorithms (rather than single and isolated input data, like CSV files). Recently a team of researchers at Free University of Amsterdam has published a paper on this topic, where the Knowledge Graph is used as default data model for Machine Learning [11].
  Alternatively, as is best known, ontologies can be useful for overcoming several natural language processing tasks [12].

Chapters 4 and 5 of this thesis refer to the extraction and representation of knowledge with a mixed approach.

To further justify this new AI trend, there are also the words of Andreas Blumauer (CEO and co-founder of Semantic Web Company and SEMANTiCS) released in an interview of 2018 [13]:
"The web as we know it has reached its limits due to a lack of semantics. It's not only about finding quickly something that is related to a search term, it's about getting informed and being able to learn from available resources efficiently. Considering all topics that are currently at the focus of public attention (such as Artificial Intelligence, Machine Learning, Data Science, Natural Language Processing, Semantic Web), in the immediate future, I cannot see anything on top of that but rather a fusion of all of that".

Chapter 2

Background on Text Mining

"Most people believe that language describes reality.
I argue that language brings forth reality."
- Matthew Budd, M.D., 2000

This chapter presents the theoretical framework that supports the thesis work in Text Mining and Natural Language Processing area.

2.1 Overview

2.1.1 Motivations

The exponential growth of the Web and its users over the past few years have contributed to increasing the addition of new sources of knowledge, by forming an enormous repository of data. Today the Internet has massive amount of text in the form of news, reviews, blogs, forums, e-mails, digital scientific libraries and especially social network documents. So, the Web can be seen as a technological enabler which encourages the creation of a large amount of text content.

The importance of Text Mining applications has consequently increased because of the high number of web-enabled applications which lead to the creation of such unstructured data (without pre-defined models that can describe them).

Unstructured textual data is the easiest form of data which can be created in any application scenario. As a result, there has been and still is a strong need to design methods and algorithms capable of process it.

IDC and Seagate predict that the Global Datashpere will grow from 33 Zettabytes in 2018 to 175 Zettabytes by 2025 and the majority of that will be unstructured [14]. Additionally, the Computer World magazine states that unstructured information might account for more than 70%–80% of all data in organizations.

It is only in recent years that awareness of the value of data has spread among companies (where only a small percentage of the information held is expressed numerically). Manual analysis of this unstructured textual data is increasingly impractical and Text Mining techniques are being developed to automate the analysis process.

If Data Mining is a field that has gained more and more interest recently [15], this is particularly true for the case of text data, where the developments that have taken place in these last months have allowed the achievement of goals deemed insurmountable until recently. So, we live in a unique time for Natural Language Processing in general and Text Mining has turned into a vital research zone.

Learning from text and natural language is one of the great challenges of Artificial Intelligence and Machine Learning. Any substantial progress in this domain has strong impact on many applications. One of the fundamental problems is to learn the *meaning* and *usage* of words in a data-driven way (possibly without further linguistic prior knowledge). The main problem a Natural Language Processing system has to address roots in the distinction between the lexical level of "what actually has been said or written" and the semantic level of "what was intended" in a text.

According to a recent Gartner report [16], Natural Language Processing and conversational analytics have a disruptive potential over the next three to five years. For the healthcare field, these subjects promise to be revolutionary.

This enthusiasm is also expressed in [17], where the authors say that *"in special domains (e.g., biomedical domain) and for special mining tasks (e.g., extraction of knowledge from the Web), Natural Language Processing techniques, especially information extraction, are playing an important role in obtaining a semantically more meaningful representation of text".*

2.1.2 Definitions

As often happens for research areas at the time of their maximum expansion (where new solutions are constantly proposed and combined with already existing ones), there is a bit of confusion from a notional point of view and clearly organizing the topics covered by the field is not immediate. This section shows the vision of the best known authors in the scientific community, who have distinguished themselves with their publications, or the definitions more consolidated in the literature.

Text Mining

Text Mining (also referred to as *Text Analytics*) is an Artificial Intelligence (AI) discipline that studies methods and algorithms for extracting significant knowledge or patterns from collections of unstructured textual data. So, its definition is one-to-one with that of Data Mining. In other words, Text Mining discloses new and already obscure data in a mechanized way.

Text Mining identifies facts, relationships and assertions that would otherwise remain buried in the mass of textual big data. Once extracted, this information is converted into a structured form that can be further analyzed, presented in a graphical form or integrated into databases, data warehouses or business intelligence dashboards.

Text Mining is not just about making statistics on text, but it includes tasks like text classification, text clustering, document summarization, sentiment analysis and any other technique that performs text processing to bring out valuable information.

Text Mining techniques are continuously applied in industry, academia, web applications, Internet and other fields [18].

In addition to having obvious intersections with Data Mining, an important characteristic of this area is that it has been explored by multiple communities such as Machine Learning, Information Retrieval, Big Data and Semantic Web [17]. Deep Learning systems have also proven to be excellent and are offering great results precisely because they work better on unstructured data such as text. As represented in Figure 2.1, Text Mining draws upon contributions of many text analytical components and knowledge input from many external disciplines. In many cases, these communities tend to have some overlap, but are largely disjoint and carry on their research independently. However, bring together studies of different communities in order to maximize the cross-disciplinary understanding of Text Mining is crucial, and one of the objectives of the thesis is precisely this.

Furthermore, it is important to underline how a distinct set of researchers have come to the fore in the field of Text Mining to face new aspects such as data streams and social networks.

Natural Language Processing

Text Mining employs a variety of methodologies to process the text and one of the most important is Natural Language Processing (NLP).

NLP is a field that covers computer understanding and manipulation of human language in a smart and useful way. It is all about leveraging tools, techniques and algorithms to process and understand natural language-based data, which is usually unstructured like text, speech and so on.

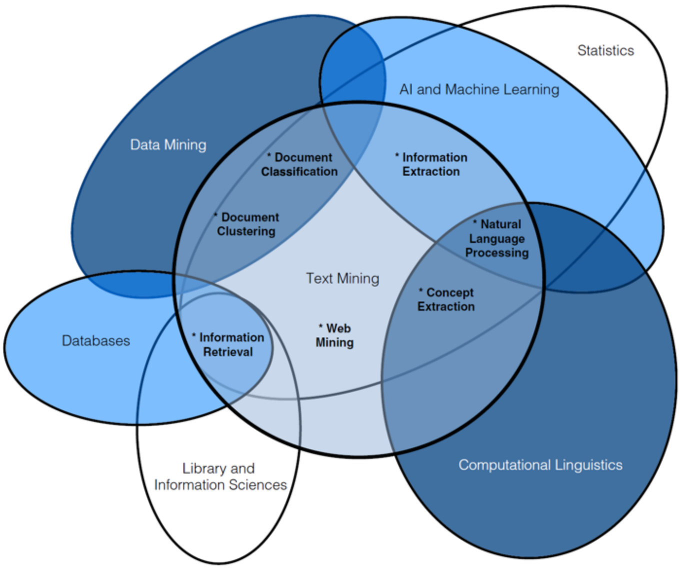


Figure 2.1: A Venn diagram of the subfields of Text Mining and how they relate

From “Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications”, 2012 [19]

It gives business people an easier way to ask questions about data and to receive an explanation of the insights.

In general, the unstructured text is easy for people, but very complex for machines. As stated in [20], difficulties with automated text comprehension are caused by the fact that the human/natural language:

- has ambiguous terms and phrases;
- often strongly relies on the context and background knowledge for defining and conveying meaning;
- gives many way to represent similar concepts;
- is based on commonsense knowledge and reasoning;
- is rarely precise;
- deals with complex social interactions.

Moreover, as evidenced by Charu C. Aggarwal and ChengXiang Zhai in [17], the most important characteristic of text data is that it is sparse and highly dimensional. Therefore, NLP and Text Mining don’t have a simple objective and are characterized as difficult problems in Computer Science.

NLP is a very broad discipline where the “Semantic Web” could also be included (if a symbolic rather than statistical approach is considered). It concerns a set of techniques where there is also "computational linguistics" (related to professional figures halfway between computer scientists and linguists, who work on computational models for text processing based on language analysis).

2.1.3 Latent Semantic Analysis

Latent Semantic Analysis (LSA) is a fully automatic mathematical/statistical technique for recognizing *latent similarities* on data, proposed by Landauer and Dumais in 1997 [21].

More specifically, it is a theory and method for extracting, inferring and representing the contextual-usage meaning of words by statistical computations applied to a large corpus of text. The underlying idea is that the aggregate of all contexts in which a given word does and does not appear provides a set of mutual constraints that largely determines the similarity between words (and groups of words) based on their meaning. Essentially, LSA induces global knowledge indirectly from local *co-occurrence* data (without using syntax, linguistic, pragmatics or perceptual information about the physical world). As stated by Landauer et al. [22], *"another way to think of this is that LSA represents the meaning of a word as a kind of average of the meaning of all the passages in which it appears, and the meaning of a passage as a kind of average of the meaning of all the words it contains"*.

Obviously LSA is general and can be applied to terms (and not just words). The method begins with the representation of the text as a traditional *term-document matrix*, where each row stands for a unique term, each column stands for a unique document, and each cell contains the frequency with which the term of its row appears in the document of its column. Therefore, given a collection of text documents $D = d_1, \dots, d_N$ with terms forming a vocabulary $W = w_1, \dots, w_M$, the data are summarized in a $N \times M$ co-occurrence table of counts $C = (n(d_i, w_j))_{ij}$, where $n(d, w) \in \mathbb{N}$ denotes how often the term w occurred in the document d.

Next, the cell entries are subjected to a preliminary transformation consisting in the application of a *weighting* function, with the aim of best representing the importance of each term in each document (e.g., tf-idf). In fact, raw counts do not work particularly well because they do not consider the significance a term has in the document in which it appears.

LSA transforms the term-document matrix by bringing out *latent semantic associations* (i.e. of a higher order and not based on lexical matching) between terms and documents. In particular, it performs a mapping of the matrix in a reduced *vector space* which approximates the original one, focusing on the essence of the data. Vectors are mathematically well suited to drawing semantic comparisons. In the transformed space (also called *"latent semantic space"* [23]),

- both terms and documents may be present;
- semantically similar or associated terms are in neighboring positions;
- the axes are no longer terms or documents, but latent variables;
- semantic similarity can be measured between any two points (terms and terms, documents and documents, terms and documents);
- the measure of similarity computed is usually, but not always, the cosine between vectors.

Figure 2.2 shows these concepts.

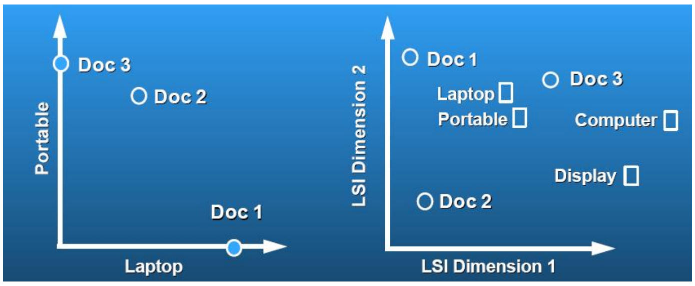


Figure 2.2: Illustration of the transition from original vector space to latent semantic space with new dimensions

Courtesy of Susan Dumais

The mapping is restricted to be linear and is based on a *Singular Value Decomposition (SVD)* of the term-document matrix. SVD is a technique in linear algebra that factorizes any matrix C into the product of three separate matrices, as follows.

$$\underset{M\times N}{C} = \underset{M\times M}{U}\ \underset{M\times N}{\Sigma}\ \underset{N\times N}{V^T} \tag{2.1}$$

U and V are two orthogonal matrices, and Σ is a diagonal matrix containing the singular values of C. Formally, $\Sigma = diag(\sigma_1, \ldots, \sigma_p)$ where $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_p \geq 0$ and $p = min(M, N)$.
Starting from this definition, the CC^T product is very interesting. $CC^T = (U\Sigma V^T)(V\Sigma^T U^T) = U\Sigma V^T V\Sigma^T U^T = U\Sigma\Sigma^T U^T = U\Sigma^2 U^T \rightarrow U\Sigma^2 U^T U = CC^T U \rightarrow U\Sigma^2 U^T U = CC^T U \rightarrow CC^T U = \Sigma^2 U$. Considering how the scalar product of two vectors is a measure of similarity, CC^T represents a square matrix where the cells contain the similarities between all terms. The columns

of U contains the right eigenvectors of CC^T, and Σ^2 contains its eigenvalues λ in descending order (consequently, $\sigma = \sqrt{\lambda}$). In the same way, C^TC is the square matrix of similarity between all documents, and the columns of V are its eigenvectors.

The singular values in Σ are the components of the new dimensions, and each of them indicates the importance of the relative dimension. Being in descending order, the first of them capture the greatest variation of the data (i.e. contain more information).

The reduction of dimensionality through SVM occurs by choosing the number of axes to be kept in the new space. In other words, SVD reduces dimensionality by selecting only the k largest singular values, and only keeping the first k columns of U and the first k rows of V^T. So, given a matrix C $M \times N$ and a positive integer k, SVD finds the matrix $C_k = U_k \Sigma_k V_k^t$ of rank at most k (between all matrices with k linearly independent vectors) that minimizes the difference with the original matrix $X = C - C_k$, according to the Frobenius norm (which is simply the traditional norm applied to a matrix).

$$\|X\|_F = \sqrt{\sum_{i=1}^{M} \sum_{j=1}^{N} X_{ij}^2} \tag{2.2}$$

Figure 2.3 resumes this process. Obviously, the matrix C_k has the same dimensions as the original one. The positions of all terms and documents in the latent semantic space are obtained respectively from the products of matrices $U_k \times \Sigma_k$ and $V_k \times \Sigma_k$.

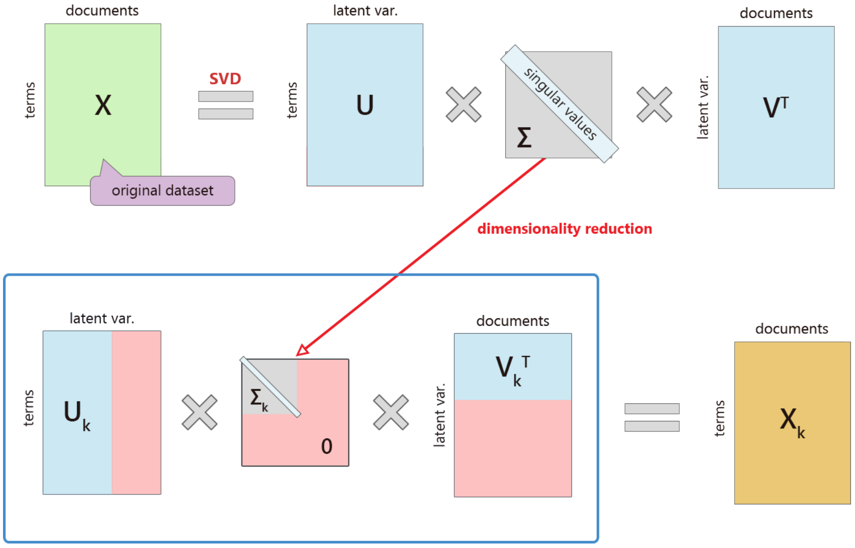


Figure 2.3: LSA dimensionality reduction through SVD

Summing up, with the decomposition created by LSA, you can bring terms and documents in a common space, whose dimensions are latent variables. This operation is called *latent semantic mapping.*

Within the reconstructed term-documents matrix C_k, the similarity between the pairs of terms or documents is measured with the cosine of the scalar product resulting from the respective similarity matrices ($C_k C_k^T = U_k \Sigma_k^2 U_k^T$ or $C_k^T C_k = (V_k \Sigma_k)(V_k \Sigma_k)^T$) and the specific instances of U_k or V_k. For example, the similarity between two documents v_i and v_j is calculated as $cos(v_i \Sigma_k, v_j \Sigma_k) = (v_i \Sigma_k)(v_j \Sigma_k)^T / (\|v_i \Sigma_k\| \, \|v_j \Sigma_k\|)$.

One of the most interesting aspects of LSA is the ability to *fold-in* new documents. In particular, for many tasks it can be useful to carry out the transposition of queries in the latent semantic space. A query q is equivalent to a set of terms in C. It must undergo the same preliminary transformations that the cell entries of C received before the SVD application. Consequently, q is first modeled in a vector form with the classic bag of words representation, and then subjected to the same term weighting scheme. So, queries can be represented as pseudo-document vectors from weighted combinations of terms. Transforming a query vector q in a new document q_k means transforming it into a row of the matrix V_k (i.e., a column of V_k^T). So, it is useful to rewrite the formula of the SVD factorization of C for V^T. $U^T C = U^T U \Sigma V^T \rightarrow U^T C = \Sigma V^T \rightarrow \Sigma^{-1} U^T C = \Sigma^{-1} \Sigma V^T \rightarrow \Sigma^{-1} U^T C = V^T$. Since $V^T = \Sigma^{-1} U^T C$, then $V = C^T U \Sigma^{-1}$. The formula of q_k derives from this observation.

$$q_k = q^T U_k \Sigma_k^{-1} \tag{2.3}$$

On the other hand, to map the documents in the latent space, the matrix V is multiplied by Σ. Therefore, to get row matrix corresponding to the position of the query in the latent semantic space we perform the product $q^T U \Sigma^{-1} \Sigma = q^T U$.

The similarity between a query q_k and a document v_i is calculated as a standard similarity between documents: $cos(q_k \Sigma_k, v_i \Sigma_k) = cos(q^T U_k, v_i \Sigma_k)$. Each $c_{ij} \in C_k$ is an association between a term u_i and a document v_j, where $x_{ij} = u_i \Sigma_k v_j = (u_i \Sigma_k^{1/2})(\Sigma_k^{1/2} v_j)$. So, the similarity between a term u_i and a document v_j is calculated as $cos(u_i \Sigma_k^{1/2}, \Sigma_k^{1/2} v_j)$.
Finally, the similarity between a term u_i and a query q is calculated as $cos(u_i \Sigma_k^{1/2}, \Sigma_k^{-1/2} U_k^T q)$.

k is a *hyperparameter* we can select and adjust to delete noise and unnecessary data, as well as better capture the mutual implications of terms and documents. Hypothetically, the optimal space for the reconstruction has the same dimensionality as the source that generates discourse, that is, the human speaker or writer's semantic space. If k is too small (the dimensionality has been reduced too much), useful data are also eliminated; if k is too large (the dimensionality has been reduced too little), noise remains. The number of dimensions retained in LSA is an empirical issue. In practice k is typically between 100 and 500 [24], but is highly dependent on the goal of the individual application. To date, there is still no way to establish it optimally (it is still being studied). Some LSA implementations, however, make heuristics available.

From the mathematical method just described, it can be seen that the similarity values estimated by LSA are not simply based on co-occurrence frequencies. They depend on a deeper statistical analysis where the information is compressed and projected on the desired space, bringing out latent semantic relationships.

Term and document meaning representations derived by LSA have been found capable of simulating a variety of human cognitive phenomena, ranging from developmental acquisition of recognition vocabulary to word-categorization, discourse comprehension, document coherence estimation, and judgments of essay quality. The correlations demonstrate close resemblance between what LSA extracts and several human abilities involving representation of meaning, association or semantic similarity. As also highlighted in the title of the paper with which LSA was presented, Laundauer and Dumais argue that its underlying mechanism can address Plato's Problem (the problem of explaining how we can know so much given our limited experience). The results have shown that the meaning similarities so derived closely match those of humans, LSA's rate of acquisition of such knowledge from text is a good approximation, and these accomplishments depend strongly on the dimensionality of the latent space. Although its representation of reality is basic, relatively simple, and surely imperfect, LSA performs a powerful and, by the human-comparison standard, correct induction of knowledge [22].

LSA applications are numerous: data clustering, document classification, topic detection (deductible from the terms that most characterize each axis), information retrieval (returning documents with supra-threshold cosine similarity to a query), synonymy (terms with reduced cosine similarity in the latent semantic space) and polysemy (terms halfway between multiple axes) handling, opinion mining and sentiment analysis, natural language understanding, etc. The main strength of LSA is to be based only on the raw text, working in an *unsupervised* way and without requiring dictionaries, knowledge bases, semantic networks, grammars, syntactic parsers, or morphologies.

However, LSA suffers from two main weaknesses:

- it has an expressiveness limited by the bag of words representation (it does not take into account the order of words and therefore cannot handle negations);

- provides a single semantic representation for the same term;

The computational cost of LSA is the same as SVD (being based entirely on it). The calculation of the exact SVD solution has a complexity a little less than quadratic, and in some applications it may take some time: $O(min\{MN^2, M^2N\})$. This performance aspect is also the reason why new queries are folded in, instead of recalculating the latent space entirely. On the other hand, the fact that LSA supports fold-ins (unlike t-SNE, for example) is a great advantage. However, there are no-exact solutions for SVD with significantly lower costs [25]. From this point of view, very efficient algorithms have recently been developed, so as to reduce calculation times by orders of magnitude compared to the exact version typically implemented in libraries. In some circumstances, this allows to obtain a reduction in execution times from a few hours to a few minutes.

Chapter 3

Background on Semantic Web

> *"The web of human-readable document is being merged with*
> *a web of machine-understandable data.*
> *The potential of the mixture of humans and machines working together and*
> *communication through the web could be immense."*
> *- Tim Berners-Lee, May 1998*

The subject of Knowledge Representation (KR) gained a new dimension with the advent of the computer age. With the evolution of the Semantic Web, KR techniques moved into the spotlight, aiming at bringing human understanding of the meaning of data to the world of machines. This chapter presents the current state-of-the-art in Semantic Web for the main purposes of this thesis, giving a detailed overview of the reasons behind this science and then detailing the concepts explored.

3.1 Overview

3.1.1 The limits of the traditional Web

The World Wide Web (WWW), commonly known as Web, has gone through several changes.

In the first part of its history, from the 1960s to the late 1980s, there was only the *Internet* and information access was difficult and for experts only.

The Web as we know it today has been developed back to 1989 by Tim Berners-Lee at CERN in Geneva, Switzerland.
Called also *Web of Documents*, it is an information system where documents and other web resources are interlinked each other and uniquely identified by Uniform Resource Locators (abbreviated URL) that specify how they can be

retrieved across the Internet from their remote location. So, while the Internet is an infrastructure, the Web is a service on top of it.
The Hypertext Markup Language (HTML) is the standard markup language used for creating web pages and web applications, which primarily contain information in natural language, digital images and multimedia resources along with the rendering instructions to be displayed for human consumption. WWW resources are instead transferred via the Hypertext Transfer Protocol (HTTP). Access to documents - which are the focus - is now simplified by the presence of web browsers with a graphical user interface.

Since its appearance on the Internet, the World Wide Web has become more and more mainstream and has grown into the world's largest repository of human knowledge. Currently, it offers a universal information sharing environment to users. Today there are not only the documents provided by some users and companies, but there is even more (especially considering social media platforms and Internet of Things). The rapid and constant growth of the amount of information on the Web has raised many research challenges such as information overloading, poor retrieval and aggregation problems. With the high amount of data available, making sense of them and finding useful information for humans are difficult tasks even for search engines which rely mostly on content-independent statistical algorithms. Syntactic variations or misspellings of the search keywords in documents prevent a reliable statistical score of document relevance.

These issues derive from the fact that the current Web is mainly designed for human consumption and not for an automated machine processing. The information stored on web pages cannot be understood by machines, while still being easily understood by people. In fact, humans have a lot of experience and contextual knowledge to determine what a web page content means. We are used to dealing with information gathering, but for machines this is a hard problem, even if we need their support to filter and to retrieve the information we are really looking for. The main point is that search engines should be able to understand the information content and to interpret it. Unfortunately HTML does not describe what the information really means. So, one of the biggest downsides of the traditional Web is that there is no explicit semantics.

3.1.2 Towards the Web of Data

The Semantic Web is an emerging research area that arises from the awareness of the importance of meaning.

Tim Berners-Lee, the Web's inventor, has coined the term *Semantic Web* (despite having said on several occasions to prefer *Data Web* as denomination [26, 27]) and in [28] provides a concise definition of it: *"The Semantic Web is* *not a separate Web but an extension of the current one, in which information is given well-defined meaning, better enabling computers and people to work in cooperation".*

Therefore, the Semantic Web (or Web of Data) really represents the third generation of the Web: an evolution of the existing Web of Documents (a "Syntactic Web") in which semantic is added to the resources, data are at the center of the processing and machines are able to comprehend them, bringing structure to the meaningful content of web pages. As a consequence, the Semantic Web can be seen as an Internet plug-in that allows web content to be understandable, interpretable, and usable not only in natural languages, but also in related software.

For the semantic web to function, computers must have access to structured collections of information and sets of inference rules that they can use to conduct automated reasoning. To realize its full potential, knowledge representation must be linked into a single global system. The challenge of the Semantic Web, therefore, is to provide a language that express both data and rules for reasoning about the data and that allow rules from any existing knowledge-representation system to be exported onto the Web [28].

Always according to Berners-Lee [26], the Semantic Web is a bit like having all databases out there as one big database (a universal network of semantic propositions). So, one of the problems the Semantic Web can solve is also helping unlock information in various silos, in different software applications, and different places that currently cannot be connected easily.

Definitely, the Semantic Web at its heart is very simple. It's just about the relationships between things. Properly designed, it can assist the evolution of human knowledge as a whole. Although the original vision dates back to 1998 [29], it is still difficult to imagine the power that comes from having so many different sorts of computer-readable and computer-understandable data within the web environment. The effectiveness of software agents based on Semantic Web will increase exponentially as more machine-readable Web content and automated services become available.

3.2 The Semantic Web Technology Stack

The Semantic Web encompasses a set of technologies with the aim of making information understandable and elaborable by computers. The goal is to create new knowledge by drawing conclusions that are implied by the initial knowledge base.

The *Semantic Web Technology Stack*, as shown in Figure 3.1, follows a layered approach. The stack is formed by the stratification of multiple components in which each level provides a set of specific functionalities and is the basis for the standards defined at the higher levels.

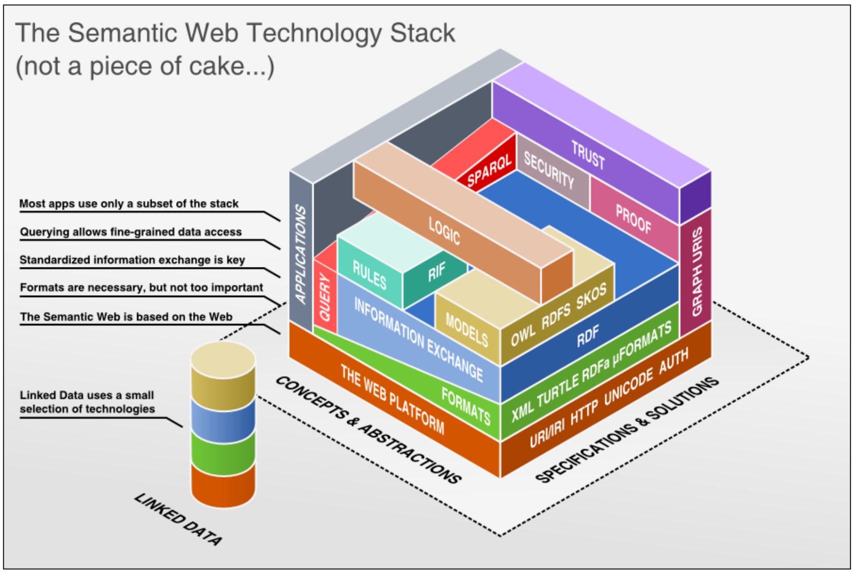

The Semantic Web Technology Stack
(not a piece of cake...)
Most apps use only a subset of the stack
Querying allows fine-grained data access
Standardized information exchange is key
Formats are necessary, but not too important
The Semantic Web is based on the Web
Linked Data uses a small selection of technologies
LINKED DATA
APPLICATIONS
TRUST
SPARQL
SECURITY
PROOF
LOGIC
GRAPH URIS
QUERY
RULES
RIF
MODELS
OWL RDFS SKOS
INFORMATION EXCHANGE
RDF
FORMATS
XML TURTLE RDFa µFORMATS
THE WEB PLATFORM
URI/IRI HTTP UNICODE AUTH
CONCEPTS & ABSTRACTIONS
SPECIFICATIONS & SOLUTIONS

Figure 3.1: The Semantic Web Technology Stack
From https://www.flickr.com/photos/_after8_/3702240268

The lowest layer consist of technologies already familiar for who knows the Web platform. These concern encoding and representation of text (UNICODE), transporting messages (HTTP), resource identification (URI) and authentication mechanisms.

On top of that, there are formats and serialization languages (like XML and Turtle).

The main building block is represented by the Resource Description Framework (abbreviated RDF), which is on the information exchange layer and describes the information contained in a web resource providing unambiguous methods to express semantics [30, 31]. RDF is the most preferred way to represent Semantic Web, and acts as the primary base for it. As suggested by the stack structure, everything encoded in RDF needs to be serialized in XML, in Turtle or in any another format before being passed in a message to another system.

Above the exchange layer there are models defined through knowledge representations like RDF Schema (various abbreviated as RDFS, RDF-S, or RDF/S) [32, 31] or Web Ontology Language (OWL) [33]. RDFS allows to define simple vocabularies used in RDF descriptions. OWL is the most popular ontology language: it is an extension of RDFS and therefore is more sophisticated and expressive.

For these models it is also possible to create logical rules or apply logical frameworks. In fact, rule languages allow writing inference rules in a standard way which can be used for reasoning in a particular domain. Among several standards of rule languages there are RuleML and the Semantic Web Rule Language (SWRL) [34]. The latter combines RuleML and OWL, and includes a high-level abstract syntax for Horn-like rules.

Moreover there is a standardised query language for RDF data, called SPARQL [35], that provides both a protocol and a language for querying RDF graphs via pattern matching. Of course there are also a security and a trust layer, which are needed to assure that the information content of resources is of high quality and can be trusted.

3.3 Resource Description Framework

In Semantic Web, data on web pages are well-defined and tagged, such that computers can directly understand, satisfy needs of users, interpret and reveal implicit knowledge.

Knowledge is represented by directed and labelled graph, where nodes are resources and edges are relationships between these resources. The *Resource Description Framework* (abbreviated RDF) is a World Wide Web Consortium (W3C) data graph model [36].

RDF is mainly intended to be used when data need to be machine processable rather than being only accessed by people. Furthermore, RDF provides a standardised way to express information such that it can be exchanged between different systems without loss of meaning [36].

RDF defines data in triple attributes, which are subjects (resources or blank nodes), predicates (resources) and objects (any element). As a result, RDF triples have the form `(subject, predicate, object)` and provide the way to make *statements* about things. Resources can be Universal Resource Identifier (URI), blank nodes and literals that are data values.
In RDF, a document makes assertions that particular things have properties with certain values. This structure turns out to be a natural way to describe the vast majority of the data processed by machines.
Subject and object are each identified by a Universal Resource Identifier (URI), just used in a link on a Web page (URLs, Uniform Resource Locators, are the most common type of URI). Predicates depict directed relationship between subjects and objects, and are also identified by URIs. The URIs ensure that concepts are not just words in a document but are tied to a unique definition that everyone can find on the Web. The Semantic Web, in naming every concept simply by URI, lets anyone express new concepts with minimal effort.

Its unifying logical language will enable these concepts to be progressively linked into a universal Web [28].

Over the past years, RDF datasets have increased and as a result have drawn the attention of more researchers. SPARQL is the query language that is used to return the necessary response from within a large RDF dataset [37]. In order to reduce response times, it is specific for RDF data. It is very important to query these data efficiently and there have been many studies related to query optimization [38, 39, 40].

3.4 Linked Open Data

The idea behind *Open Data* (shortened as OD) is closely similar to the concept of open source software [41]. According to the Open Definition the essence of Open Data can be summed up in the statement: *"Open means anyone can freely access, use, modify, and share for any purpose (subject, at most, to requirements that preserve provenance and openness)"* [42].

Open Data refers to the publication of any collection of data in a machine-readable format, with no licensing or patent restriction so that everyone is free to use, reuse and redistribute for any purpose. Governments provide transparency and increase increase public participation through Open Data. Scientific institutions can benefit from Open Data for deriving new knowledge and insights, as well as sharing it. Entrepreneurs can use the data to support their business and decision making.

Strictly related to the concept of Open Data is the concept of *Linked Data* (shortened as LD). Linked Data refers to *"data published on the Web in such a way that is machine-readable, its meaning is explicitly defined, it is linked to other external datasets, and it can in turn be linked to from external datasets"* [43].

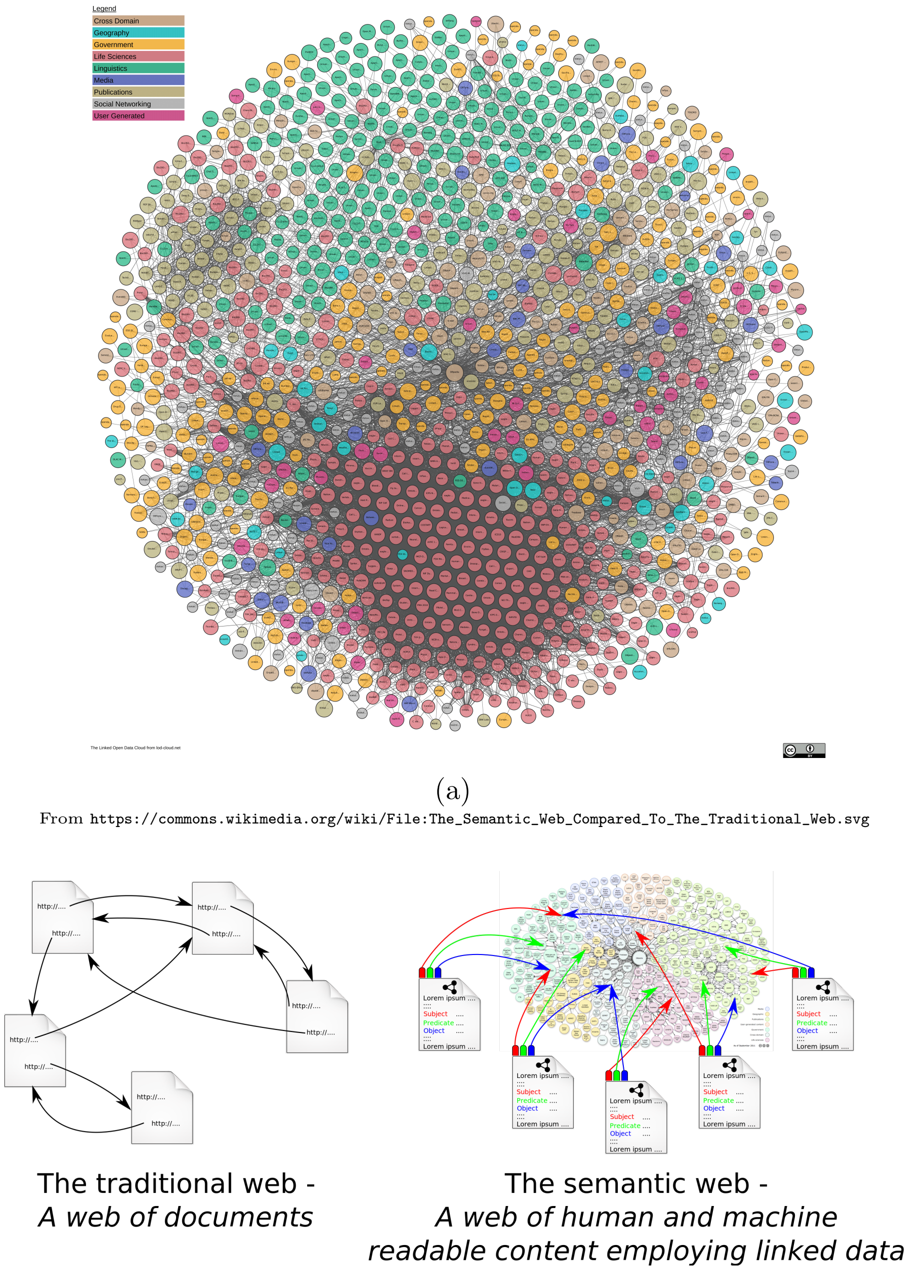
The merger of the movement of Open Data with the concept of Linked Data gives raise to a powerful data organisation and knowledge distribution. The *Linked Open Data* (abbreviated LOD) - as the combination of Open Data and Linked Data - is the set of publicly available (RDF) data on the Web, accessible via HTTP, identified via URI and linked to other data in the same way. LOD constitutes a huge graph and is also an official W3C project.

An independent community page for Linked Data [44] describes it as *"using the Web to connect related data that wasn't previously linked, or using the Web to lower the barriers to linking data currently linked using other methods"*.

In the Web of Data there are right now about 1.250 datasets and 16.000 links among each other, with DBpedia[1] at the center of this cloud [44]. Figure

[1] The semantic version of Wikipedia. `https://wiki.dbpedia.org/`

3.2 offers a graphical representation of the actual LOD cloud diagram and the differences with the Web of Documents.



(a)

From `https://commons.wikimedia.org/wiki/File:The_Semantic_Web_Compared_To_The_Traditional_Web.svg`

(b)

From `https://commons.wikimedia.org/wiki/File:The_Semantic_Web_Compared_To_The_Traditional_Web.svg`

Figure 3.2: The Linked Open Data Cloud (March 2019) and the comparison with the traditional Web

Many projects (like BBC music) now use publicly available linked data sources in order to avoid the costs and complexity of updating and maintaining data. With LOD:

- information is dynamically aggregated from external, publicly available data;
- usually there is no need of screen scraping;
- no specialized APIs are required;
- data access is via simple HTTP Request;
- data is always up-to-date without manual interaction.

However, it must also be said that some of the LOD datasets, despite being publicly available, have poor quality, discouraging others from contributing to and reusing them. Some emerging researches are therefore seeking to add value to Linked Open Data [45].

3.5 Ontologies

A formal model is needed to represent knowledge (especially in a mechanised way) and make it understandable by machines. One way to achieve that is to use ontologies.

Ontology itself is a concept which was already known in antiquity. The word *Ontology* comes from the Greek *ontos* (being) and *logos* (study) and has its root in philosophy where it refers to the theories about the nature of existence, the reality, as well as the basic categories of being and their relations [46]. In other words, the term Ontology is used to mean *"the study of categories of things that exist or may exist in some domain"* [47].

Even though there is no universal definition for ontology, one of the most frequently cited in Computer Science literature comes from Thomas R. Gruber [48]: *"an ontology is an* ***explicit***, ***formal*** *specification of a* ***shared conceptualization***. *The term is borrowed from philosophy, where an Ontology is a systematic account of Existence. For AI systems, what 'exists' is that which can be represented"*. Here, *conceptualization* stands for an abstract model of the world within the domain considered (identified relevant concepts and relations among them). The conceptualization must be *shared* because it has to capture consensual knowledge (i.e., accepted by a group and not only by an individual); *explicit* in the sense that the meaning of all concepts must be defined (there should be nothing left undefined) and *formal* to indicate that the ontology must be machine readable and understandable.

Ontologies provide a flexible and expressive tool for modelling high-level concepts and relations among them, which allows both system and the user to operate the same taxonomy [49]. So, an ontology is a formal description of knowledge as a set of concepts within a domain and the relationships that hold between them. To enable such a description, we need to formally specify components like classes, individuals and relations as well as restrictions, rules and axioms. Classes are abstract groups, sets, or collections of objects representing ontology concepts. Classes are also characterised via attributes (name-value pairs) and can be related to other classes. So, relations are special attributes, whose values are objects of (other) classes. For relations and attributes, rules (constraints) can be defined to determine allowed values. Classes, relations, and constraints can be put together to form complex statements. Formal axioms are a special case of statements and represent knowledge that can't be expressed simply with the help of other existing components. Ontologies can be represented as a hierarchically structured set of concepts describing a specific domain of knowledge. Therefore, one of the benefits of using them is to be able to provide a common ground that can lead to a shared understanding for the same concepts.

In a knowledge base, instances of classes describe individuals and usually are distinguished by classes themselves: classes and relationship among them form the so-called terminological knowledge (TBox), while instances constitutes the assertional knowledge (ABox).

As a result, ontologies do not only introduce a sharable and reusable knowledge representation but can also add new knowledge about the domain [50].

Ontologies provide users with the necessary structure to link one piece of information to other pieces of information on the Web of Open Linked Data. Ontologies also enables database interoperability, cross-database search and smooth knowledge management. They are particularly useful when you want to reuse contents and features, supporting the idea that there is no need to reinvent the wheel [51].

Ultimately it can be said that ontologies function like a “brain”. They “work and reason” with concepts and relationships in ways that are close to the way humans perceive interlinked concepts [50].

3.5.1 Differences from other knowledge descriptors

Ontologies are not the only method that use formal specifications for knowledge representation. For example, there are also taxonomies and thesauri (that always belong to the world of Semantic Web). However, unlike the latter, ontologies express relationships and enable users to link multiple concepts to other concepts in a variety of ways. Moreover, in an ontology the types of relationships are greater in number and more specific in their function. Even if they don't appear on the Semantic Web Technology Stack, this section offers a brief distinction between the concepts of controlled vocabularies, taxonomies and thesauri [52, 53].

Controlled Vocabularies

A Controlled Vocabulary (CV) [54] is an approach for the representation of concepts, and not just words. It is an authoritative list of terms to be used in indexing (human or automated). It does not necessarily have any structure or relationships between terms within the list. It is "controlled" because only terms from the list may be used for the covered subject area. Controlled vocabularies often handle synonyms. They can be viewed as a broadest category which includes taxonomies, thesauri and ontologies (that are consequentially specific kinds of controlled vocabularies).

Taxonomies

A taxonomy is a controlled vocabulary with a hierarchical structure, where terms have relations to other terms within the taxonomy itself but attributes are not permitted. Taxonomies use generic *father-son* relationships in order to enable the definition of different types of hierarchies. More specific relations are, for example, *part-of* and *cause-effect*. Based on the above, taxonomies are often displayed as a tree structure where terms are nodes.

Thesauri

A thesaurus is a more structured type of controlled vocabulary that provides information about each term and its semantic relationships to other terms within the same thesaurus. According to the standards, the types of thesauri relationships fall into three categories: hierarchical, associative and equivalence. If other relationships than those supported are required, one must resort to more general ontologies. In addition, it is common in thesauri that some or all terms have scope notes, short explanations of how the term should be used in indexing.

A notable example of thesaurus is *WordNet*. WordNet is a large lexical database of semantic relations between words in more than 200 languages [55], developed at the Princeton University under the direction of George A. Miller [56]. WordNet links include synonyms, hyponyms, and meronyms. Nouns, verbs, adjectives, and adverbs are organised into set of synonyms (called synsets), each representing a lexicalised concept with a short definition and usage examples.

WordNet can thus be seen as a combination and extension of a dictionary and thesaurus. While it is accessible to human users via a web browser, its primary use is in automatic text analysis and artificial intelligence applications. It also forms the basis of several other projects, such as - for example - MultiWordNet [57] and SentiWordNet [58].

3.5.2 Ontologies Classification

In the literature, various different ontology classifications have been proposed over time. As depicted in Figure 3.3, Guarino et al. proposed a classification based on the degree of generalisation and conceptualization subject (scope) [59]:

- *Top Level Ontologies*: describe very generic and abstract concepts such as space, time, event, insubstantial or concrete objects, etc. Ontologies of this kind are independent to the domain usage.

- *Domain Ontologies*: describe a given domain (e.g., medicine or university) by specialising the concepts provided by the Top Level Ontology, independently to the task.

- *Task Ontologies*: describe the vocabulary of terms needed to perform generic tasks or activities (e.g., diagnosis) by specialising the concepts provided by the Top Level Ontology.

- *Application Ontologies*: describe the terms of concepts depending both on a particular domain and task. Ontologies of this kind are restricted only to a specific application.

Modularization can be used at each level. For instance, Top Level Ontologies could be imported by ontologies at lower levels and adding them specific knowledge. Domain and Task Ontologies may be independent and are combined to be used for Application Ontology.

There also different classifications suggested by other authors like the ones in [60, 61].

3.5.3 Automatic Reasoning

Inference

Inferences are steps in reasoning, moving from premises to logical consequences; etymologically, the word infer means to “carry forward”.

Inference is theoretically traditionally divided into deduction, induction and abduction, a distinction that in Europe dates at least to Aristotle.

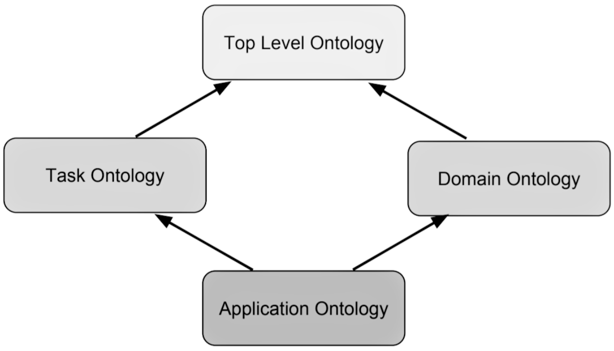


Figure 3.3: Guarino's Ontology classification where thick arrows represent specialisation relationships

Image adapted from [59]

Deduction is inference deriving logical conclusions from premises known or assumed to be true. *Induction* is inference from particular premises to a universal conclusion. *Abduction* starts with an observation or a set of observations and then seeks to find the simplest and most likely explanation for the observations.

Without inference, is required to pay the cost of finding, encoding, storing and maintaining an explicit statement of every possible needed fact.

Inference can also be used to identify inconsistencies, and this is particularly useful during data integration.

Reasoning

Reasoning is the process of extracting new knowledge (inferring facts that have not been explicitly stated) from an ontology and its instance base. It is one of the most powerful features of Semantic Web technologies.

A software agent without the ability to make inference would only be able to answer questions related to syntax (and not semantics). So, semantics is a prerequisite for reasoning support.

A *Semantic Reasoner* (also known as *reasoner engine* or simply *reasoner*) is a software system whose primary goal is to infer knowledge which is *implicitly* stated by reasoning upon the knowledge *explicitly* stated, according to the rules that have been define. As anticipated, the reasoners are also used to *validate* an ontology, making sure that it does not contain any inconsistencies among its term definitions.

Description Logics

A machine that uses a First Order Logic (shortened as FOL) for knowledge representation and reasoning is based on a deductive process. However the deduction mechanism is a semi-decision procedure: when the conclusion cannot be deduced from the premises, the procedure may not end. So, FOL has intractability and semi-decidability characteristics that do not make it adequate. For this reason the logical formalism for ontologies is usually provided by Description Logics (DLs).

DLs are a subset of the first-order predicate logic for which efficient reasoning support is possible, and a family of formal knowledge representation languages widespread in the field of artificial intelligence [62].

DLs sacrifice expressiveness in exchange for security regarding the decidability of the reasoning. Therefore, a DL is a compromise between expressive power and reasoning complexity, where the logic remains fairly expressive for the applications, the deduction procedure ends in any case after a finite number of steps and has an acceptable computational cost.

Ontologies formalized with Description Logics are a powerful tool for capturing and reasoning on complex hierarchical dependencies between concepts used in domain of interest [63].

A lot of research is currently being focused on investigating this compromise as well as the discovery of efficient reasoning algorithms applicable to practical situations [64].

There are different types of reasoners and each one is more or less suitable depending on the characteristics that are considered. Three classical open source reasoners available are: HermiT [65], Racer [66], Pellet [67] and FaCT++ [68]. Some surveys report an exhaustive comparison between them [69].

3.5.4 Ontology Languages

Ontology languages allow users to write explicit, formal conceptualizations of domain models. The main requirement are:

- *a well-defined syntax*, that is a necessary condition for machine-processing of information;
- *a formal semantics*, which describes the meaning of knowledge precisely and allow people (and computers) to reason about the knowledge itself;
- *an efficient reasoning support*, in order to make derivations mechanically (consistency, unintended relationships between classes, classification etc.);
- *a sufficient expressive power.*

The richer the language is, the more difficult the reasoning support becomes. Although complex language constructs allow to represent more knowledge, computation becomes inefficient and eventually undecidable (sometimes crossing the border of *noncomputability*). When it comes to choosing an ontology language for the Semantic Web there is always a trade-off between expressive power and efficient reasoning, depending on the kind of application to be designed.

3.5.5 OWL

RDFS is deliberately intended to be a light ontological language for the definition of simple vocabularies, but in many cases to address the demands of the Semantic Web more expressiveness is needed. In different contexts it is also necessary to model property constraints and their characteristics.

The Web Ontology Language (abbreviated OWL) [70] is the most popular ontology language. It is so the *de facto* standard for publishing and sharing ontologies in the Semantic Web, as well as a W3C Recommendation.

In other words, OWL is a semantic web computational logic-based language, designed to represent rich and complex knowledge about things and the relations between them. It extends RDFS to overcome its limitations and it also provides detailed, consistent and meaningful distinctions between classes, properties and relationships.

OWL enriches ontology modeling in semantic graph databases, also known as RDF triplestores. Used together with an OWL reasoner in such triplestores, it enables consistency and satisfiability checks.

OWL introduces many new language primitives which extend RDF and RDFS, such as union, intersection and complement operators. Also, OWL comes equipped with means for defining equivalence and difference between instances, classes and properties. As a consequence, in OWL is possible to state that two classes are disjoint or are the same (despite being identified with different URIs). These relationships help users match concepts and ensure the disambiguation between different instances that share the same names or descriptions.

It is also possible to use restrictions on properties such as *cardinality* or specify that a certain property is *transitive*, *symmetric* and *reflexive.*

Like RDF Schema, OWL can be serialised using RDF syntax. Furthermore, it adopts the *open world assumption* (which means that missing information is treated as unknown) and the *not unique name assumption* (different identifiers may refer to same entities in the real world).

The use of OWL for the explicit modeling of relationships between classes, combined with inference through descriptive logical reasoning, enables the management of a poly hierachy and the creation of powerful queries.

OWL consists of a family of three languages with different degrees of expressivity and computational properties: OWL Full, OWL DL and OWL Lite.
OWL Full is the most expressive OWL language and uses all the OWL language primitives. This flexibility comes at the expense of decidability. In fact, all of the typical reasoning tasks over an OWL Full ontology are undecidable.
OWL DL and OWL lite are two restricted forms of the OWL language, aimed at ensuring decidability and at allowing an efficient reasoning support.

While ontologies provide a rich set of tools for modeling data, their usability therefore comes with certain limitations. One such limitation concerns the availability of property constructs. For example, while providing powerful class constructs, the most recent version of the Web Ontology Language (OWL2) has a somewhat limited set of property constructs.

3.6 Uncertainty and Ontologies

Today uncertainty is a fundamental issue. Being able to manage it is becoming a prerogative of more and more contexts and applications.

The web content is, for the most part, subject to imperfection. But, as this thesis wants to underline, it is necessary to adopt a broader point of view than just the Semantic Web because this is indispensable in complex projects that require the integration of different research areas (like the case study presented in Chapter 6). In fact, also the knowledge extracted with Text Mining corresponds to statistical theories. For example the truth value of the rules obtainable from a decision tree is probabilistic and not deterministic. In the same way, an increasing number of tasks addressed through Machine Learning and Deep Learning techniques accompany their results with probabilistic values in terms of accuracy and confidence. The same observation can be applied to many other fields, also external to Computer Science. In complex open-world problems, legislating unambiguous usage is often infeasible.

Ignoring this kind of data during knowledge representation is a mistake and a serious approximation. To think still today that knowledge is exclusively deterministic is anachronistic. Therefore, a statistical knowledge representation model is needed.

Outside the Semantic Web, a recent work published in the literature concerns the union of Probabilistic Logic Programming (abbreviated PLP) and neural networks [71].

Although ontologies have become the standard for representing knowledge on the Semantic Web, they have a primary limitation: the inability to represent vague and imprecise knowledge. The majority of ontology modeling formalisms and tools are based on crisp logic and do not provide ways to represent uncertainty. So, standard ontologies are crisp[2]. Classical ontology languages are not appropriate to deal with uncertainty in knowledge, which is inherent in most of the real world application domains. As a result, ontology description languages become less suitable in all those domains in which the concepts that must be represented have not a precise definition. If we take into account that we have to deal with the example scenarios mentioned above, then it is easily verified that this problem is, unfortunately, likely the rule rather than an exception. Even more so if we consider the representation of knowledge extracted directly from unstructured text.

The need to deal with vague information in Semantic Web languages is rising in importance and, thus, calls for a standard way to represent such information. Much research has been undertaken to extend ontologies with the means to overcome this limit and has resulted in numerous extensions from crisp ontologies to fuzzy ontologies (that can allow user question answering with probabilistic values). As stated in [73], *"it is necessary and desirable to do the best we can with the knowledge we have. To pretend certainty when we are uncertain is not doing the best we can. To do this in a sound and principled manner requires a sound and principled way to represent, communicate, and reason with uncertainty".* In 2011 a W3C standard for a fuzzy OWL extension was not expected in the near future [74] and today the situation has not changed.

The original web ontology language and tools were not designed to handle fuzzy and probabilistic information. The lack of support for uncertainty in OWL is a serious limitation for a language expected to support applications in uncertainty-laden domains such as biogenetics or medicine.

Augmenting an ontology to carry numerical and/or structural information about fuzzy/probabilistic relationships is not enough to deem it a fuzzy/probabilistic ontology [73]. In general, people faced with the complex challenge of representing uncertainty in languages like OWL tend to begin by writing probabilities as annotations (e.g. marked-up text describing some details related to a specific object or property). This is a palliative solution that addresses only part of the information that needs to be represented for the lack of a good representational scheme that captures structural constraints and dependencies among probabilities. A true probabilistic ontology must be capable of properly representing those nuances.

[2] A crisp ontology is a precise (i.e., binary) specification of a conceptualization [72]

Therefore, additional research has focused on two different approaches. In fact, this issue can be addressed extending current Semantic Web languages to cope with vagueness, or by providing a procedure to represent such information within current solutions (for example, using a concrete methodology based on OWL 2 annotation properties). Regarding this second choice, several fuzzy extensions have been proposed over time in the literature [75, 76, 77] (sometimes considered insufficiently expressive [74]) and a recent review of the state-of-the-art can be find in [78].

In general, there is a widespread opinion that for adequate representation of uncertainty in OWL some language extension is necessary, be it at the level of language syntax or of higher level patterns [79]. An example of the first is Fuzzy RDF [80], which transforms the RDF model from the standard triple to the couple (value, triple), adding to the triple a certainty value indication.

As asserted in [72], *"the problem to deal with imprecise concepts has been addressed several decades ago by Zadeh [81], who gave bird in the meanwhile to the so-called fuzzy set and fuzzy logic theory. Unfortunately, despite the popularity of fuzzy set theory, relative little work has been carried out in extending ontology description languages towards the representation of imprecise concepts".*

Some of the proposed models lack formal semantics and are sometimes unclear [82, 83]. Others discuss the development of a Fuzzy OWL and a Fuzzy DL, while originally relying on classic logics based knowledge representation [84, 85]. Still others try to reduce a fuzzy ontology to a crisp one [86, 87], or are based on different paradigms but are prototypical and or not adequately supported by tools.

In the next sections the two major detected solutions (equipped with a tool) for fuzzy and probabilistic ontologies representation and reasoning are presented. Both have been considered suitable for the objectives of the thesis.

3.6.1 Fuzzy OWL 2

As can be deduced from the name, Fuzzy OWL 2 [74] is a fuzzy extension for OWL 2 where fuzzy syntactic elements (like fuzzy ABox, datatypes, modifiers and weighted sum concepts) are encoded using annotation properties in order to allow the use of existing OWL 2 editors and reasoners (which automatically discard, where possible, the fuzzy part of a fuzzy ontology and producing the same results as if it would not exist). Therefore, it is based on *fuzzy description logics* and a simple syntactic extension to OWL. It defines a framework - designed to be easily extensible - that represent fuzzy ontologies using current languages and resources.

The authors have implemented a Protégé plug-in [88] to edit fuzzy ontologies according to their extensional approach and some parsers that translate them into the languages supported by fuzzy DL reasoners.

The suggested methodology is the following:

- creation of the core part of the standard ontology (also supporting standard reasoners) by using any OWL 2 ontology editor;
- introduction of the fuzzy part of the ontology by using annotation properties (eventually with the help of the available Protégé plug-in and its graphical interface);
- parser-based translation from OWL 2 with annotations into languages supported by some fuzzy DL reasoners (such as fuzzyDL [89] and DeLorean [90]).

3.6.2 PR-OWL 2

PR-OWL [73, 91] is a probabilistic extension to OWL. It belongs to a different family of approaches focused on modeling *probability* and provides a consistent framework for building probabilistic ontologies. So, probabilistic ontologies provide a means to represent and reason with uncertainty. They are used for the purpose of comprehensively describing knowledge about a domain and the uncertainty associated with that knowledge in a principled, structured and shareable way, ideally in a format that can be read and processed by a computer.

The team behind PR-OWL argue that the ontology layer is the appropriate place in the Semantic Web architecture (described in Section 3.2) for representing declarative knowledge about likelihood.

PR-OWL is based on Multi-Entity Bayesian Networks (MEBN), a first-order Bayesian logic that integrates classical first-order logic with probability theory (chosen as best trade-off between flexibility and expressiveness among the candidate probabilistic logics).

Classical first-order logic (FOL) is by far the most commonly used, studied and implemented logical system, serving as the logical basis for most current generation AI systems and ontology languages. MEBN logic provides the basis for extending the capability of these systems by introducing a logically coherent representation for uncertainty. Because a MEBN theory represents a coherent probability distribution, Bayes Theorem provides a mathematical foundation for learning and inference, that reduces to classical logic in the case of certain knowledge (i.e., all probabilities are zero or one).

The OWL compatibility has been significantly improved with the release of the second version of PR-OWL from Carvalho: PR-OWL 2 [92].

The key principle of PR-OWL 2 to representing fuzzy information is the separation of crisp ontology from fuzzy information ontology. While an ontology can be built from scratch using standard tools, it is also possible to extend an existing crisp ontology (base ontology) to represent fuzzy information.

The suggested methodology is similar to that of Fuzzy OWL 2 and is the following:

- creation of the core part of the standard ontology (also supporting standard reasoners) by using any OWL editor;
- loading of the OWL file inside UnBBayes tool with PR-OWL plug-in [93];
- import of the set of classes, subclasses and properties provided by PR-OWL;
- definition of the upper probabilistic ontology using PR-OWL definitions to represent uncertainty about their attributes and relationships.

From that definition it is possible to realize that nothing prevents a probabilistic ontology from being "partially probabilistic". It is the knowledge engineer who chooses which concepts to insert in the "probabilistic part" of the ontology, while writing the others in standard OWL.

3.7 Knowledge Graphs

Given the thesis topics, the concept of Knowledge Graph (KG) plays a very important role. In this section, this technology is presented with reference to the available scientific literature and the differences with ontologies are illustrated.

3.7.1 Google Knowledge Graph

Starting in 2012, Google began developing a so-called Knowledge Graph [94, 95] with the aim of providing more meaningful answers within automated dialogue systems. In the original post with which Google's Amit Singhal introduced this technology, the Knowledge Graph is described as *"a system that collects information about real world objects and their connections: a critical first step towards building the next generation of search"*. The blog post, now become famous, therefore announces a radical change in the methodology behind Google's long-stand mission to organize the world's information and make it universally accessible and useful. In fact, Knowledge Graph represents a totally different approach than the Google's original PageRank algorithm where links and other contextual signals were used to determine the relevancy of a webpage. Google's attention shifts for the first time from "strings" to "things", potentially realizing Tim Berners-Lee's early notion of a Semantic Web (described in Chapter 3) where all content is linked according to its meaning. Since then a kind of hype has arisen around the term "Knowledge Graph". However, there are very few technical details about Google Knowledge Graph (GKG) organization, coverage and size (of which it is only known that at the time of the announcement it contained 500 million objects and more than 3.5 billion facts about relationships). There are also very limited means for using this resource outside Google's own projects. Users and software agents can interact with it, but its API returns only individual matching entities rather than graphs of interconnected entities. In its early days, the Knowledge Graph was partially based on *Freebase*[3], a famous general-purpose knowledge base that Google acquired in 2010. Today, the Knowledge Graph still uses *schema.org*, a collaborative effort between multiple tech giants to develop a schema for tagging content online [96]. On the surface, information from the Knowledge Graph is used to augment search results and to enhance its AI when answering direct spoken questions in Google Assistant and Google Home voice queries. Behind the scenes and in return, Google uses its KG to improve its machine learning.

3.7.2 Current Definition

Knowledge Graph does not yet have a well-established definition.
It represents a collect of interlinked descriptions of entities, where descriptions have a formal structure and each entity represents part of the description of the entities related to it (forming a network).
As stated in [97], knowledge graphs combine characteristics of several data management paradigms and can be understood as database (for structured queries support), graph (for network data structure) and knowledge base (for formal semantics representation). However, a knowledge graph is not like any other database; it is supposed to provide new insights, which can be used to infer new things about the world.
Knowledge Graph technology means being able to connect different types of data in meaningful ways. So, a knowledge graph is necessarily built on semantics. Knowledge graphs have the ability to project information into a multidimensional *conceptual space* where similarity measures along different dimensions can be used to group together related concepts. This allows for an

[3] `https://developers.google.com/freebase`

integrated solution that not only identifies the meanings of entities, clustering them into a unified knowledge layer, but also correlates concepts to allow for inference generation and insights [98].

3.7.3 Differences with ontologies

As stated in [96], *"it's unlikely that a consensus will emerge anytime soon on what a knowledge graph is or how it is different from an ontology"*. From a terminological point of view, there is still a strong debate in the literature [99, 100, 101, 102] and a consensus on how exactly this technology differs from *ontologies* has not yet been achieved. Many agree that knowledge graphs resemble ontologies but they are not exactly the same thing.

They are both represented by nodes and customized semantic relationships, appearing potentially the same in visualizations. Moreover, they are both based on RDF triples (comprising subject-predicate-object), and are usually also built on the Semantic Web standard OWL. A knowledge graph may comprise multiple domain ontologies, vocabularies or knowledge information system, as well as an ontology can import other resources.

According to the authors of [103], a knowledge graph can be considered ontologies and more. It can be said that a knowledge graph is a way an ontology could be represented. In the community many claim that the real divide between ontology and knowledge graph has nothing to do with size or semantics, but rather the very nature of the data. The real point is that knowledge graphs are fact-oriented, while ontologies are schema-oriented. Unlike knowledge graphs, in domain ontologies the focus is not on data (or facts), but on a highly expressive description (disjointness, cardinality, restrictions) of the concepts, the relations between them and useful annotations (synonyms, definitions, comments, design choices, etc.). An ontology is metadata/schema, whereas the knowledge graph is the data itself. An ontology usually deals with concepts, not instances of concepts. So, not every RDF graph is a knowledge graph.

The commonly accepted difference is therefore subtle. In the world of Linked Open Data, for this reason they are often used interchangeably (just think of DBPedia). By supporting both the definition of classes and instances (as well as properties), OWL can be used to represent both ontologies and knowledge graphs. Ontologies are generally regarded as smaller collections of assertions that are hand-curated, usually for solving a domain-specific problem. By comparison, knowledge graphs can include literally billions of assertions, just as often domain-specific as they are cross-domain.

Apart from disagreements on size, role and separation of classes from instance data, the two approaches are fundamentally the same. In summary, they are generally larger or smaller versions of each other, with more or less sophisticated knowledge encoding techniques.

3.7.4 Applications and famous examples

Knowledge graphs are all around. Today Google, Facebook and Microsoft use them as part of their infrastructure.

Enterprises use knowledge graphs in order to combine disparate data silos, bring together structured and unstructured data, provide a unified view of varied unconnected data sources, answer complex queries, and make better decisions by finding things faster.

Knowledge graph-based solutions are generally used to improve dialog-based access to information. More generally, knowledge graphs can serve various roles and provide many benefits. They support search, recommendation engines, e-commerce, and enterprise knowledge management. From a point of view related to Text Analysis (central for this thesis), knowledge graphs provide background knowledge to enable a more accurate interpretation of text. Furthermore, they also enable the return of results accompanied by semantic tags. Facts extracted from the text can be added to enrich the knowledge graph.

Those graphs are often constructed from semi-structured knowledge, such as Wikipedia, or harvested from the web with a combination of statistical and linguistic methods. The result are large-scale knowledge graphs that try to make a good trade-off between completeness and correctness.

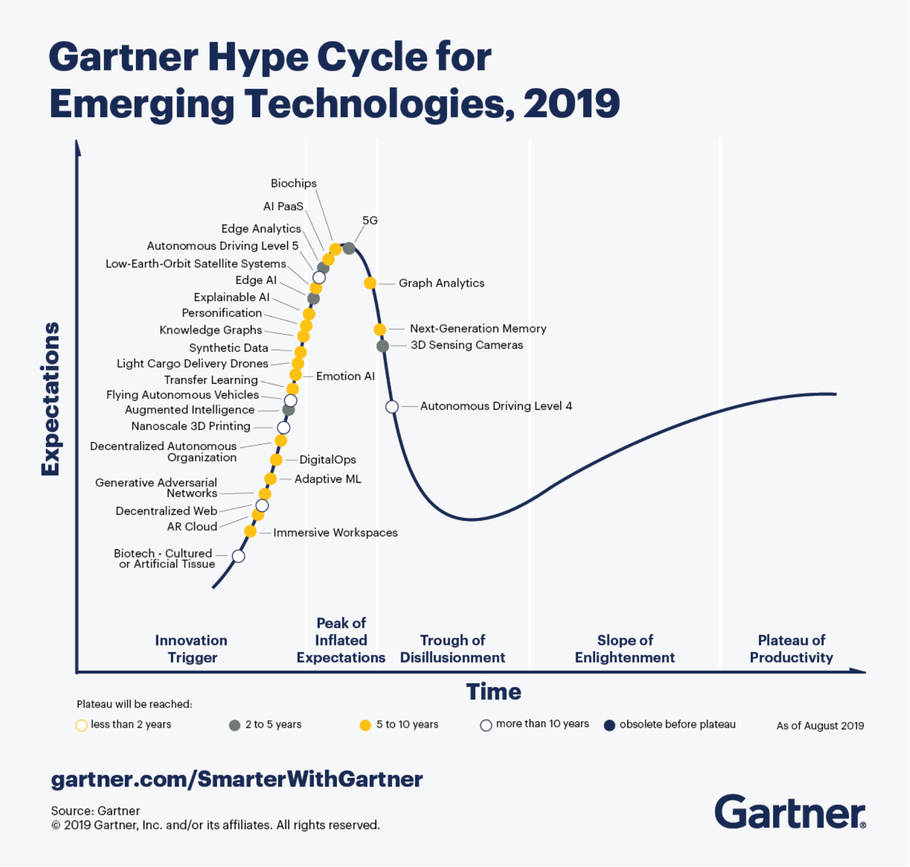
In addition to *Google Knowledge Graph*, the most famous examples of knowledge graphs are *Freebase* [104], *DBpedia* [105], *YAGO* [106], *GeoNames* and *Wordnet* [107].

3.7.5 Trend

As shown in Figure 3.4, the Gartner[4] Hype Cycle for Emerging Technologies (2019) [108] considers Knowledge Graphs as one of the Innovation Trigger with the highest expectations, preventing a peak in this sense between 5 and 10 years.

Always Gartner, the previous year affirmed *"the rising role of content and context for delivering insights with AI technologies, as well as recent knowledge graph offerings for AI applications have pulled knowledge graphs to the surface"*.

[4] The world's leading research and advisory company



Source: Gartner
© 2019 Gartner, Inc. and/or its affiliates. All rights reserved.

Figure 3.4: Gartner Hype Cycle for Emerging Technologies, 2019
From [108]

3.8 Semantic Web for Health

Since the birth of this research area, many Semantic Web applications have been destined for the healthcare sector.

It is interesting to note how the use of the Semantic Web to facilitate the discovery of treatments for diseases was already discussed in the Q&A by Tim Berners-Lee in 2007 [26].

Now that we have ontologies from an expert user view for genomics data, proteomics data, epidemiological data, phenotypes and so on (overcoming the problems related to keeping data in different departments and different pieces of software), we have reached the time to introduce the patients view (main objective of this thesis).

The enormous and always increasing amount of social data is the ideal key to achieve that purpose. The Semantic Web is a technology designed to specifically do that, to open up the boundaries between the silos, to allow scientists to explore hypothesis, to look at how things connect in new combinations that have never before dreamt of.

3.8.1 Rare Disease Ontologies and Resources

This section lists the main resources available today for the representation of rare diseases (and diseases in general) in health information systems. Many ontologies that have been done on diseases can be found in BioPortal web page.

Disease Ontology

The Disease Ontology (DO) [109] represents a comprehensive knowledge base of human diseases. It is hosted at the Institute for Genome Sciences at the University of Maryland School Of Medicine.

The project was initially developed in 2003 at Northwestern University.

Today has the purpose of providing the biomedical community with consistent, reusable and sustainable descriptions of human disease terms, phenotype characteristics and related medical vocabulary disease concepts. So, it has a nosology use.

The Disease Ontology semantically integrates disease and medical vocabularies through extensive cross mapping of DO terms to MeSH, ICD, NCI's thesaurus, SNOMED and OMIM [110, 111].

Medical Subject Headings

The Medical Subject Headings (MeSH) [112] is a biomedical vocabulary used to index MEDLINE and other databases. It contains subject headings (canonical terms) organized into an eleven-level hierarchy and 83 subheadings. It is typically used to label documents or to classify medical content in general. For example, the psychiatric disorder schizophrenia is assigned the MeSH term "*Schizophrenia*" and organized into the hierarchy "*Psychiatry and Psychology → Mental Disorders → Schizophrenia and Disorders with Psychotic Features → Schizophrenia*".

The MeSH vocabulary also includes a vast number of entry terms, which are intended to be synonyms of the canonical heading terms. It also supports other features like conceptual relations.

MeSH is used to annotate articles on PubMed.

Finally, MeSH descriptors are divided into 16 categories, and the titles within each category can be more or less specific.

Orphanet

Orphanet [1, 113] it is considered one of the most important references for rare diseases domain and is so used by various applications.

Born in France, it is the result of a European project (with a network of 40 countries worldwide) and represents a qualified reference source for high quality information on rare diseases and orphan drugs, with the aim of ensuring equal access to knowledge for all stakeholders. In more detail, it serves as a reference portal since 1997 for RDs populated by literature curation and validated by international experts. In other words, Orphanet responds to one of the major needs in health information systems and for research: share and/or integrate data coming from heterogeneous sources with diverse reference terminologies.

Orphanet information system is supported by a multilingual relational database designed around the concept of "disorder" (unlike OMIM which instead defines entries on their genetic basis). Moreover, Orphanet offers services adapted to the needs of patients and their families, health professionals, and researchers.

The database integrates the nosology (or classification) of rare diseases, their relationship with genes and epidemiological data, cross-references to other terminologies, expert centres, diagnostic tests and professionals. To date more than 6.000 disorders are registered and new data is added regularly.

It also develops a nomenclature for rare diseases (called "ORPHA Number"), which is essential to increase the visibility of these disorders in health and research information systems.

Orphanet Rare Disease Ontology

After the launch of the multi-hierarchical classification in 2008 on the occasion of the fourth version of Orphanet, researchers (both academics and industry) showed a strong interest in the use of this dataset and pushed the authors of Orphanet itself to move on to an ontological approach for knowledge representation (seeing in the Semantic Web a powerful means for their work).

The Orphanet Rare Disease Ontology (shortened as ORDO) [114, 115] is an open-access OWL ontology derived from the Orphanet information system and updated twice a year with periodic extractions from it (starting from XML files related to the partial exports of the Orphanet database itself). It was jointly developed in 2013 by Orphanet and the European Bioinformatics Institute (EMBL-EBI) in order to allow for the association of new content and the establishment of new research hypothesis.

ORDO is recommended as the primary "disease nomenclature" ontology in many projects, often playing a central role (similarly to DBpedia in the LOD graph).

It also provides disease cross references to the International Classification of Diseases (ICD-10), SNOMED Clinical Terms (SNOMEDCT), Medical Subject Headings (MeSH), Medical Dictionary for Regulatory Activities (MedDRA), Online Mendelian Inheritance in Man (OMIM) and Unified Medical Language System (UMLS) and genes are cross-referenced to HUGO Gene Nomenclature Committee (HGNC), Universal Protein Resource (UniProt), Ensembl, Reactome and Genatlas.

ORDO enables complex queries about rare diseases, their epidemiological data (age of onset, prevalence, mode of inheritance) and gene-disorder functional relationships.

ORDO 2.9 (the latest version at the time of writing) consists of 14.559 classes and 205.428 annotation assertion axioms. Each concept from the Orphanet database forms a distinct OWL class and is associated with other classes using a set of defined object properties. The ontology contains an hierarchical clinical classification of rare disorders, each described by its preferred name and synonyms. In addition, each rare disorder entry is assigned to a unique phenome type (such as “disease”, “malformation syndrome”, “clinical syndrome” and “morphological anomaly”). In ORDO a clinical entity is either a group of rare disorders, a rare disorder or a subtype of disorder.

ORDO also represents the relationship between the disorders and their genetic cause (if known), and not just the nosology such as that captured in the Disease Ontology.

The main OWL classes modeled by ORDO are shown in the Figure 3.5.

The use of class descriptions in OWL enable more complex querying which was difficult or impossible using the existing relational database.

Human Phenotype Ontology

The Human Phenotype Ontology (HPO) provides a structured and controlled vocabulary for the phenotypic features (e.g. disease symptoms) encountered in human hereditary and other diseases. HPO itself does not describe individual disease entities but, rather, the phenotypic abnormalities associated with them [115].

It provides comprehensive bioinformatic resources for the analysis of human diseases and phenotypes.

HPO was initially published in 2008 with the goal of enabling the integration of phenotype information across scientific fields and databases. Since then, the project has grown in terms of coverage and has also become part of the core Orphanet rare disease database content. In fact, the Orphanet inventory of rare disorders is enriched with annotations about HPO.

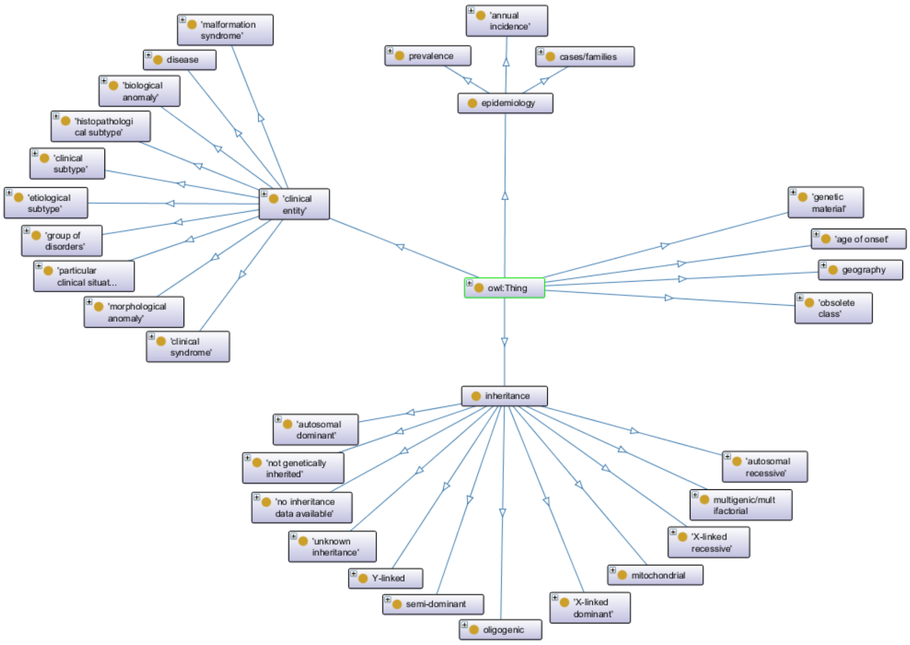


Figure 3.5: ORDO main class hierarchy
Obtained with OntoGraph Class View tool in Protégé

The “computable” descriptions of human disease using HPO phenotypic profiles have become a key element in a number of algorithms being used to support genomic discovery and diagnostics. Phenotype ontologies allow to standardize signs, symptoms, classifications and complete clinical phenotypes. They are also helpful resources for checking associations between symptoms and laboratory data. In addition, phenotype ontologies allow interoperability between registries and other resources, such as biobanks or omics databases.

Considering the general framework of this thesis, it is interesting to observe how many HPO annotations have been deduced by carrying out analyzes with Text Mining techniques on a PubMed corpus dated 2014. These annotations were validated against a manually curated subset of disorders and experimental results showed an overall precision of 67%. The motivation related to this process is that the overwhelming majority of clinical descriptions in the medical literature are available only as natural language text. HPO authors themselves say that collecting phenotypic patient data from text is not trivial but necessary [116].
Always with this in mind, it is interesting to underline how the authors also remember that clinical terminology is often unfamiliar to patients. The HPO consortium has therefore increased the usability of the HPO by patients, as well as scientists and clinicians, by systematically adding new, plain language

terms, either as synonyms to existing classes or by tagging existing HPO class labels as “layperson”.

HPO & ORDO Ontological Module

The Orphanet Rare Disease Ontology (ORDO) and the Human Phenotype Ontology (HPO) are considered the most relevant ontologies in the rare disease research. While ORDO is mainly used for naming diseases (e.g., “idiopatich achalasia”), HPO is used for describing the clinical phenotype observed in a patient (e.g., “muscle weakness”).

The combination of HPO together with Orphanet has always been considered a promising resource for automated rare disease classification. The ORDO and HPO developers have recently worked on the integration of both ontologies, annotating Orphanet’s phenome types with appropriate HPO terms. The result of this effort for interoperability is called HPO & ORDO Ontological Module (HOOM) [117] and its first version was released on March 8, 2018.

The “Clinical Entity” branch of ORDO has been refactored as a logical import of HPO, and the HPO-ORDO phenotype disease-annotations have been provided in a series of triples in OBAN format[5].

HOOM is provided as an OWL file and offers extra possibilities for researchers, pharmaceutical companies and others wishing to co-analyse rare and common disease phenotype associations.

Others

As we have been able to guess from the previous sections, the number of ontologies, vocabularies and databases in the health sector is really very high. Even if we have limited ourselves to highlighting the main resources for rare diseases and those most commonly used within projects (more than 300 biomedical ontologies embrace basic, translational and clinical science), it is important to also mention the following (some already anticipated in the description of others for obvious reasons of interoperability). These are not closely related to the world of the Semantic Web but it is important to know them to better appreciate the content of the ontologies available today.

- **Online Mendelian Inheritance in Man**
  OMIM [118] is a comprehensive, authoritative compendium of human genes and genetic phenotypes that is freely available and updated daily.

[5] Model used to express the frequency and the provenance of associations, in an ontological design context

- **SNOMED CT**
  Is a systematically organized computer processable collection of medical terms providing codes, terms, synonyms and definitions used in clinical documentation and reporting. SNOMED CT is considered to be the most comprehensive, multilingual clinical healthcare terminology in the world.

- **International Classification of Diseases**
  The International Classification of Diseases (ICD) is is a classification system (recognized as international standard) in which diseases are sorted into groups on the basis of defined criteria. The major difference with SNOMED CT is that ICD is a classification which is limited to disease.

- **Medical Dictionary for Regulatory Activities**
  MedDRA Is a clinically validated international medical terminology dictionary (and thesaurus) used by regulatory authorities in the pharmaceutical industry.

- **Unified Medical Language System**
  The UMLS is mainly a metathesaurus to enable interoperability between computer systems.

3.8.2 Interesting Applications

When one wonders how such ontologies can be used, numerous application examples can be found in the literature. Some of the most representative applications are briefly described below.

In [119] is described the use of HPO for the semantic unification of common rare diseases. A graph is constructed for modeling phenotypic similarities between diseases. The synonyms available in HPO are used to make Concept Recognition within PubMed abstracts. TF-IDF is used to identify the phenotypes of greatest interest for each disease.

In [120] HPO is used for a differential diagnosis process with semantic similarity searches. The p-value calculation is used for the construction of a ranking of candidate diseases, starting from the specification of a patient's symptoms.

In [121] ORDO is integrated with the Radiology Gamuts Ontology (RGO).

In [12] an ontology containing words considered to be of epidemiological relevance is used for disease name extraction from Twitter messages in order to boost a neural architecture.

3.8.3 Considerations

As stated in a recent survey about the state-of-the-art of Semantic Web for Healthcare [122], there are certain limitations and challenges on the current use of Semantic Web technologies in this sector (which continues to produce large volumes of heterogeneous data).

The efforts to ontology development in this domain are quite fragmented and non-standardized.

If it is true that a single ontology is not enough to describe the various data, it is equally true that several competitive ontologies exist to describe thing from the same domain. So, there is also an overlapping problem that does not facilitate the Linked Data objective.

Mappings from proprietary formats to ontology concepts are difficult and intensive task. Maintenance of ontologies and datasets is another challenge and there are not enough tools sufficiently advanced to facilitate the overcoming of these problems.

Moreover, despite the large number of ontologies for modeling medical data, there is still lack of resources considering also a non-clinician perspective. To the best of the research carried out for this work, there are no ontologies modeled directly from the direct opinions of patients. Considering the incredible impact that such a solution could have on those who live with a rare disease (and not only), the investigation of an approach patient-centered and the making available of a first solution to the problem is one of the main contributions of this thesis.

Chapter 4

A novel Knowledge Explanation-based Method for Descriptive Text Mining

"Knowledge has to be improved, challenged, and increased constantly, or it vanishes."
- Peter F. Drucker, Management consultant, educator and author

"Some men see things as they are and say why.
I dream things that never where and say why not."
- Robert Francis Kennedy, American politician, New York Senator

Though the knowledge extracted from mass of data can improve human capabilities in developing intelligent systems, it is still very difficult to understand and explain the knowledge models generated by the most powerful modern learning approaches, such as deep neural networks. Explaining the learned knowledge is essential to gain trust from experts who need to validate before using it, particularly in the medical sector. Recent accidents with self-driving car and adversarial attacks that can make neural networks look stupid have highlighted the urgent need of new methods for understanding the learned knowledge. Existing solutions are still preliminary, often based on very partial explanations, highly supervised and with several limitations, such as the frequent need for large labelled datasets to achieve satisfactory results. In this chapter, we propose a novel method of descriptive text mining capable of offering accurate probabilistic explanations in unsupervised settings.

4.1 Introduction

The amount of textual data is continuously growing. This leads to a demand for newer and more efficient data analytics procedures, also able to better capture semantics. At the research level, text mining sees the presence of numerous contributions on *predictive* analysis methods, while much less exists in literature for *descriptive* ones. Predictive analytics and descriptive analytics are the two major classes with which text mining techniques can be divided according to their objectives, exactly as it happens in data mining [123]. Together they are often called "Knowledge Discovery in Data" or KDD, although the name is more suitable for descriptive analytics only. Predictive analytics is centered around statistical models and it is used to estimate the likelihood of a future outcome based on the available data. It includes text classification and sentiment analysis tasks, employing also machine learning approaches. A typical example that fall into this category is the training of algorithms for the opinion classification of reviews as positive, negative or neutral. Descriptive analytics is instead focused on the explanation of interesting phenomena from historical data, discovering the precious reasons behind a success or a failure in the past. An example in this case is understanding *why* people have made positive or negative judgments. So the role of descriptive analytics is often traced back to dashboards or simpler tasks (such as word frequencies, strongest correlations among words and topic extraction), when in reality it refers to a larger and more complex problem. Another aspect that typically distinguishes predictive from descriptive analyzes is the fact that predictive ones usually require supervision and large labeled datasets, whereas for descriptive investigations this is not always necessary.

Making descriptive text mining therefore means explaining knowledge, and *explainability* today is a very important and requested property. Highly advanced methods such as deep neural networks (DNNs) are not yet able to explain what they have learned in a human-understandable form, limiting the effectiveness of these systems. Deep learning algorithms have achieved considerable performance among many NLP tasks like POS tagging [124], sentiment analysis [125, 126], syntactic analysis [127], and machine translation [128]. However, these successful models are usually applied in a black-box manner because no information is provided about what exactly leads them to arrive at their predictions [129]. Recent works demonstrate how simple it is to deceive such networks with adversary inputs, and all this further increases the need to investigate the thought process behind them and their reliability [130, 131, 132, 133]. As autonomous machines and black-box algorithms begin making decisions previously entrusted to humans (increasing in complexity and demonstrating spoofability in certain cases), it becomes necessary for these mechanisms to explain themselves. Explainable Artificial Intelligence (XAI) [134, 135, 136] aims to solve this lack of transparency, but only recently we are witnessing embryonic and prototype early works. Furthermore, understanding a model's classification decision represents an even greater challenge and an interesting research opportunity in text mining, due to the high dimensionality in the feature space [137]. In medicine, above all, where wrong decisions by a system can be harmful, the development of explainable models is considered essential [136].

As stated in [138], there is not much work about how to generate model-unaware explanations globally in a model-agnostic way. Most of the existing solutions are therefore based on local explanations and have several limitations. Moreover, approaches that lead to the best results often require a large amount of training data (difficult or expensive to produce in many domains).

In this thesis we propose a novel method of descriptive text mining, capable of offering explanations about whole behaviors (and not just on individual instances). Furthermore, the general approach of the contribution is completely based on unlabeled data and works so with unsupervised settings.

The presented method allows to explain the learned knowledge and consequently it can be applied to a multitude of various tasks. For instance, it can be used to understand the reasons behind an average hotel review score, similarly to the aspect-based sentiment analysis (which can be seen as a specific case of descriptive text mining) but without the need of establish anything a priori. It can also be used to understand the main causes of fatal plane crashes directly from their reports. It can be employed to comprehend the factors that led to the success of some social posts. As an additional example, it can be adopted in the field of rare diseases to understand what are the aspects that lead patients to speak ill of a certain experimental medical treatment.

4.2 Related Work

4.2.1 Descriptive Text Mining

Descriptive text mining is a topic that is poorly covered in literature. The description of phenomena from past data is often treated as a set of sub-tasks, each aimed at highlighting a certain type of useful information and exploring a particular aspect. The same vision is shared by many online tools, such as IBM Watson Analytics [139]. Given the explosion of health care data in the last decade and the need for data analytical expertise, the medical domain is often taken as a reference to verify the potential of these techniques [140].
The focus of descriptive analytics is frequently placed on pattern discovery. Consequently, some works also apply general data mining methods to text (as

sequential data). In an interesting but dated paper, Ahonen et al. [141] propose the use of episodes rules and episodes (a modification of association rules and frequent sets) with textual data to discover phrases, co-occurring terms and interesting regularities. The approach shows reasonable results. It can be used to meet different objectives depending on preprocessing, but weighting is not always discriminating enough. In general, association rules are not sufficiently expressive for the broader purposes of descriptive analysis.

4.2.2 Aspect-Based Sentiment Analysis

In its basic version, sentiment analysis is considered a classification problem. However, more and more applications require a level of detail that a document or phrase level classification of opinionated text cannot provide. As shown in Figure 4.1, an overall rating should be a summary of the opinions addressed to the individual attributes of the entity itself (such as cleanliness, food, quietness and kindness for a hotel). Consequently, a positive opinionated document about a particular entity does not mean that the author has positive opinions on all aspects of the entity, and vice versa. Aspect-based sentiment analysis (ABSA) is to all intents and purposes a specific case of a more general task which is descriptive text mining. In both cases, it is crucial to build trustworthy explainable text classification models that are capable of explicitly generating fine-grained information for explaining their predictions. One of the most critical sub-tasks of aspect-based sentiment analysis is the “aspect expression extraction”. In many applications, the selection of the aspects to be evaluated is carried out manually by an expert user, who specifies them in a supervised manner. A list of the main unsupervised methods is instead reported in the survey of Liu and Zhang [142]. In general, four main lines of work can be identified.

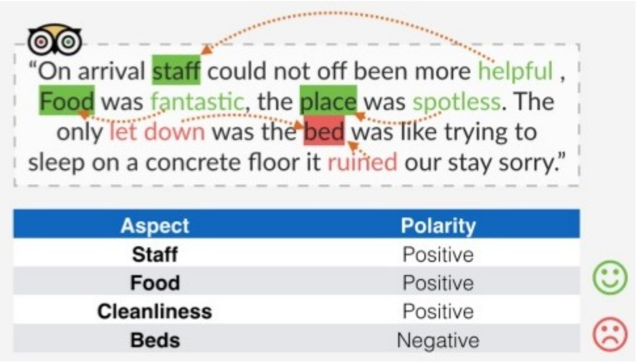
- *Frequent nouns and noun phrases identification* [143, 144].
  Since multiple words can refer to the same aspects, this kind of approaches still requires human evaluation. In addition, it is based on the assumption that infrequent nouns are likely to be non-aspects or less important aspects. Frequent nouns not corresponding to entity aspects can be found with low metrics scores regarding meronymy discriminators (like “has”, “of”, “comes with”).

- *Relationships with opinion words* [145].
  Nouns in conjunction with opinion words in a sentence can be extracted as aspects. For example, from sentence “this treatment is incredible” it can be deduced that “treatment” is an aspect.

- *Topic modeling* [146, 147, 148, 149, 150].
  Kind of solutions typically based on Latent Dirichlet Allocation (LDA) [151] or its extensions.

- *Associations between aspects and opinion/sentiment ratings* [152, 153, 154, 155].
  Statistical methods aimed at discovering aspects from documents expressing opinions.

It is worth nothing that the primary focus of the work presented in this chapter (even if independent of the sentiment analysis task) concerns the approaches supported by the fourth group of the methods listed above.



"On arrival staff could not off been more helpful , Food was fantastic, the place was spotless. The only let down was the bed was like trying to sleep on a concrete floor it ruined our stay sorry."

Aspect	Polarity
Staff	Positive
Food	Positive
Cleanliness	Positive
Beds	Negative

Figure 4.1: Example of Aspect-Based Sentiment Analysis

Obtained from "A Hierarchical Model of Reviews for Aspect-based Sentiment Analysis, Sebastian Ruder, 2016"

4.2.3 Foundational Concepts in Explainable AI

Explainable Artificial Intelligence (XAI) is a young and rapidly growing research area that aims to better understand the behavior of algorithms. A great historical background on the reasons that led to the birth of this field is reported in [156]. According to Samek, Wiegand, and Müller [157], there are four main reasons why we need explainability: trust from users (to be sure about the correctness of the decisions), modification of the system (to facilitate debugging and improve performance), learn from the system (to extract knowledge and strategies unexpected by human), and moral and legal issues (to solve problems such as assigning responsibility in case of wrong decisions).

Explainability is closely related to the concept of interpretability. Although many articles consider explainability and interpretability as two interchangeable terms, some authors believe that there are important reasons to distinguish between them [135]. In the latter case, interpretability is often seen as a property with which an explanation can be assessed, together with completeness (i.e., the accuracy of an operation description from a mathematical point of view). Doshi-Velez and Kim [158] consider the interpretability as the ability to explain or to present the decision-making process of a model in understandable terms to a human. Another elegant definition of interpretability from Miller [159] states that interpretability is the degree to which a human can understand the cause of a decision.

4.2.4 Explanation Methods Taxonomy

In recent years several methods have been proposed to deal with the problem of explainability, in particular for deep neural networks. Given the broad set of existing approaches for achieving varying degrees of interpretability and completeness, different taxonomies have also been outlined. An interesting categorization of the solutions available in the literature is defined by Gilpin et al. [135]. Considering the way the explanation is returned, they identify three major classes.

- *Processing.* Explanations that, while admittedly non-representative of the underlying decision processes, provide some degree of justification for emitted choices.
- *Representation.* Explanations that provide insights about the internal operation of a network.
- *Explanation Producing.* Networks that are specifically built to explain themselves and designed to simplify the interpretation of an opaque subsystem.

Another useful observation for an overall picture, before briefly describing the main techniques usually adopted, is that of Chhatwal et al. [137]. In reviewing the work previously done in the explainable text classification, the authors describe the various approaches they have identified: *intrinsically explainable models* (such as decision trees and If-Then rules), and *approximated explanations for complex models not directly understandable by humans* (such as deep learning models).

As evidenced in [138, 156], there are two other major distinctions made in XAI research. Firstly, there are both general explanation methods applicable to multiple models (*model-unaware explanations*) and specific explanation methods that can analyze the construction of a specific model in depth (*model-aware explanations*). Secondly, there are both explanation methods for specific instances (*local explanation methods*) and explanation methods for the whole input space (*global explanation methods*).

4.2.5 Explanation Methods Review

Existing approaches for explainable machine learning systems are mainly built around the interpretation of the outputs or the connections between inputs and outputs.

Local Interpretable Model-Agnostic Explanations (LIME) is a robust and model-unaware framework that can be applied to a wide range of ML and DL algorithms. Proposed by Ribeiro et al. [160], it uses a simple linear model (*proxy model*) to imitate the behavior of the complex one, on a region around a single instance. It can highlight the features in an image or a text which are responsible for a prediction. Therefore, this method has the benefit of yielding simple explanations, such as whether a particular word in a document or a shape in a photo is driving the model's predictions. More in detail, it attempts to determine the most salient features behind any decision by feeding through the black box model inputs similar to the original ones (inputs perturbation) and observing how the predictions change. It has also disadvantages like inconsistency among explanations of different samples. Ribeiro et al. also make *global* explanations by presenting a set of representative local explanations to users one at a time. Obviously this is not a real solution to the problem, since the construction of an overall vision is left to the user himself. This method is easy to fail when there is too much training data, and users cannot remember a large number of local explanations.

Samek et al. [157] directly built connections between the inputs and the outputs, using heatmap analysis to visualize how much each hidden element contributes to the predicted results.

Other methods of explaining complex models, such as Deep Learning Important FeaTures (DeepLIFT) [161], peer into the inner workings of a neural network to extract the most crucial parameters for the model's output. SHapley Additive exPlanations (SHAP) [162], a recent addition to this set of methods, unifies these prior attempts at interpreting model output.

Landthaler et al. [163] discussed an explanation approach similar to LIME, called eXplainable Semantic Text Matching (XSTM), for an *unsupervised* machine learning pipeline. This method is aimed at solving the Semantic Text Matching (STM) problem, or the identification of implicit logical or semantic relationships between fragments of text. So it can be used to understand the behavior of similarity measures between text with TFIDF and word embeddings. In fact, while word embeddings' characteristics are intriguing, to date it is not understood why certain structures occur in the embedding spaces.

Waltl et al. [164] investigated the application of LIME to explain the behavior of supervised machine learning algorithms on classification tasks.

Previous work on explainability for NLP was primarily focused on text classification problems. Recent research found that a prediction-based approach is often used to identify snippets of text (usually ranging from 50 to 200 words, and called "rationales") as an explanation for the classification of a document.

Zaidan et al. [165] proposed a machine learning method to use annotated rationales (text snippets and their corresponding classes) to boost text classification performance. Using SVM and annotate rationales as training data, their experiments show an important performance improvement over the baseline SVM variants.

Martens et al. [166] described a method in which the explanation of a document classification is a minimal set of the most relevant words, such that removing all the words in the set from the document itself would change the classification result.

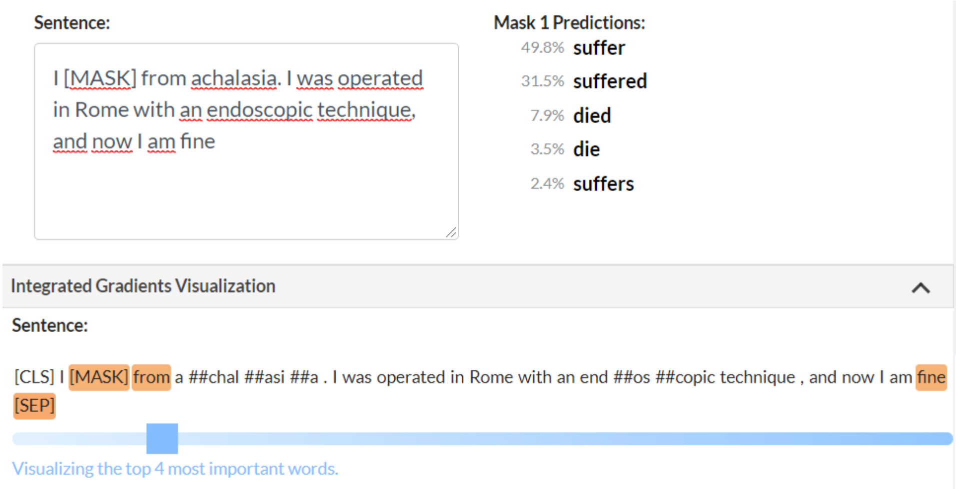
Zhang et al. [167] presented a Convolutional Neural Network (CNN) model where the explanation for the predictions is given by a probability value associated with each sentence, indicating the support of the document-level classification.

Lei et al. [168] introduced a new solution for explainable text classification. Their approach combined two components trained to operate together: a rationale generator and a rational encoder (in order to decide the class to associate with each text snippet candidate). This explanations methods outperformed the baseline on multiaspect sentiment analysis and was also successfully applied to a question retrieval task.

Although LIME method and other approaches are promising and mathematically reasonable, many papers criticize their inability to generate explanations in natural forms. In fact, fine-grained information (e.g., textual explanations for the labels) is often not considered and the systems do not explicitly produce explanations in a human-readable way. Some works therefore try to overcome this problem. In [169], Liu et al. introduced a model-agnostic generative explanation framework for text classification. However this approach is supervised and, although it can theoretically be applied to multiple problems, it is highly class dependent and requires a great effort in building a suitable training set.

In [137], Chhatwal et al. proposed a rationale generation method consisting of two phases and starting from a set of training documents tagged with a responsive label. In the first phase, a traditional classification model is generated using the training set and is deployed to identify the potentially responsive documents. In the second phase, the same model is applied to generate one or more rationales for each of the identified responsive documents, assigning also a probability score between 0 and 1. The interesting aspect of the work concerns the determination of the optimal size of the text snippets. The authors adopt an iterative approach, breaking a document into relatively large snippets and continuing to break those with large probability scores into smaller sizes. This process continues until probability scores stop increasing. The experiments show much better results considering also manually highlighted snippets in the training set.

Wallace et al. [170] recently presented AllenNLP Interpret[1], an opensource, extensible toolkit built on top of AllenNLP [171] for interpreting NLP models. Considering the importance of this framework in the scientific community, it is interesting to note the solutions chosen by the authors to have flexibility. AllenNLP Interpret considers two types of instance-level interpretations: *gradient-based saliency maps* (similarly to [129]) and adversarial attacks, which can be *HotFlip* [172] or *Input Reduction* [173]. HotFlip replaces words to change the model's prediction; Input Reduction removes word to maintain the model's prediction (like [166]). Figure 4.2 and Figure 4.3 provides two interesting examples.



Sentence:

I [MASK] from achalasia. I was operated in Rome with an endoscopic technique, and now I am fine

Mask 1 Predictions:

49.8%	suffer
31.5%	suffered
7.9%	died
3.5%	die
2.4%	suffers

Integrated Gradients Visualization

Sentence:

[CLS] I [MASK] from a ##chal ##asi ##a . I was operated in Rome with an end ##os ##copic technique , and now I am fine [SEP]

Visualizing the top 4 most important words.

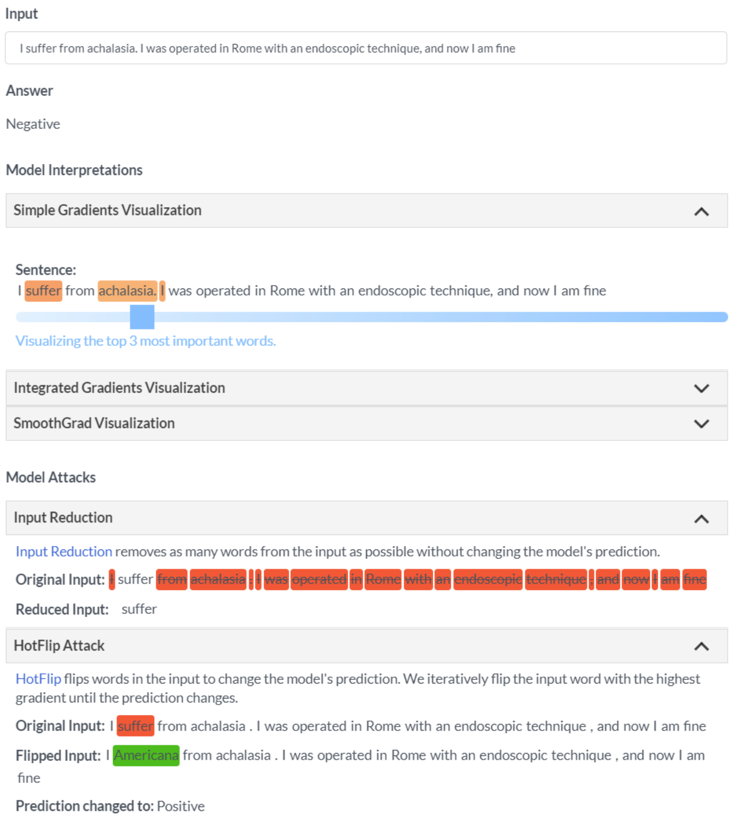
Figure 4.2: Gradient-based interpretation for BERT model on Masked Language Modeling task, using AllenNLP Interpret. The interpretation shows that BERT uses the terms "from" and "fine" to predict "suffer".
Obtained with `https://demo.allennlp.org/`

In summary, many open-source implementations exist for instance-level interpretation methods, but in literature global explainable approaches are almost totally absent. In addition, many solutions are model- or task-specific (e.g., sentiment analysis). A large number of labeled data is often required for

[1] See `https://allennlp.org/interpret`

good results. In conclusion, although there are already numerous works, much research is still needed in this area.

The contribution presented in this dissertation concerns a global unsupervised knowledge explanation method for descriptive text mining, without a strict dependency to a specific model or task.



Input

I suffer from achalasia. I was operated in Rome with an endoscopic technique, and now I am fine

Answer

Negative

Model Interpretations

Simple Gradients Visualization

Sentence:
I suffer from achalasia. I was operated in Rome with an endoscopic technique, and now I am fine

Visualizing the top 3 most important words.

Integrated Gradients Visualization

SmoothGrad Visualization

Model Attacks

Input Reduction

Input Reduction removes as many words from the input as possible without changing the model's prediction.

Original Input: ~~I~~ suffer ~~from achalasia . I was operated in Rome with an endoscopic technique , and now I am fine~~

Reduced Input: suffer

HotFlip Attack

HotFlip flips words in the input to change the model's prediction. We iteratively flip the input word with the highest gradient until the prediction changes.

Original Input: I suffer from achalasia . I was operated in Rome with an endoscopic technique , and now I am fine

Flipped Input: I Americana from achalasia . I was operated in Rome with an endoscopic technique , and now I am fine

Prediction changed to: Positive

Figure 4.3: Gradient-based and adversarial attacks interpretation for a LSTM model with GloVe embeddings on a binary sentiment classification (positive or negative), using AllenNLP Interpret. The interpretation shows that "suffer" is the term responsible for the predicted negative class, also highlighting the weakness of the model operating at document-level (which does not consider the word "fine").

Obtained with `https://demo.allennlp.org/`

4.3 A novel Descriptive Text Mining Method

The definition of an original method of descriptive text mining is the core contribution of this thesis. The whole process ranges from the preprocessing stage till the evaluation of the final results. The overall approach comprises several phases, some fundamental and some useful for a more complete analysis. In the following, all the phases are discussed in detail, paying attention to the importance of their combination in the context of the presented method. The work is described in three steps with the aim of providing a clear exposure of the contents. Section 4.3.1 illustrates the method in its interactive version (where the user is part of the learning process). Section 4.3.2 shows the transition to a fully automatic version. Finally, Section 4.3.3 presents some observations on the level of modularity, reflecting on how some phases can be pursued with a different technique (also based on the application cases) without altering the structure of the method itself. Some final considerations are instead reported in 4.3.4.

4.3.1 Interactive knowledge extraction

The various phases are listed below in chronological order.

Named Entity Recognition

Nowadays more and more advanced NER systems are available, capable of supporting numerous domains, pre-defined categories, and languages. Some of them also return information such as the identifier of the entity detected within Wikipedia and other knowledge bases. Furthermore, the limit of having only one type associated with an entity has also recently been crossed. Under this point of view, the choice of an ontology is necessary for the organization of the categories, overcoming the sub-problem of names classification according to the type of entities to which they refer. If the first works in literature focused on flat and relatively small type systems, the most recent publications address very fine-grained types organized in a hierarchical taxonomy [174, 175, 176]. Modern solutions are based on neural models such as machine learning. Their goal is to develop practical and domain-independent techniques in order to detect named entities with high accuracy automatically.

The union of all these advances in terms of research make the labeling of entities recognized by a NER within documents a phase of great importance. The use of tools with pre-trained NERs (also on specific domains) allows the unsupervised categorization of terms contained in the corpus of documents (e.g., places, foods, symptoms). This gives the opportunity for more in-depth analyzes of entity types, a better understanding of the description for the phenomenon of interest, and the possibility of being able to connect these concepts to those of already existing knowledge bases (thanks to any identifiers).

Its placement as a first step within the method is justified by the typical dependence of NER systems on preprocessing operations, which could make them ineffective. For example, both spaCy and Stanford NER models rely on letter casing for identifying named entities [177]. Most of the time the Named Entity Recognizers are designed to be applied directly on the untreated text, and any preprocessing operations required by a specific implementation (e.g., part-of-speech, lemmatization) are carried out directly behind the scenes. Furthermore, since this type of computation requires some time to be performed on a large number of documents, it is advisable to think of saving the results and reusing them in multiple analyzes.

Documents filtering by date

Applying a filter to select only the documents in a date range of interest for analysis is fundamental. If temporal information is available in the starting dataset, this stage enables the possibility of extracting descriptions of a phenomenon in specific time windows (task often required).

Quality preprocessing

Unstructured text typically has numerous imperfections (such as grammatical and spelling errors), and various types of noise. Documents from the same corpus can also have different encodings. These problems are particularly frequent in the case of sources such as blogs and social networks. The stage of quality preprocessing concerns the application of a sequence of operations aimed at transforming the textual content of documents in order to increase their quality. By cleaning up the text (e.g., encoding uniformization, symbols normalization, word lengthening fixing), it promotes the identification of the concepts and relations between them, improving the results of all subsequent phases. In fact, the quality of the data provided as input to a model determines the quality of the results themselves. Furthermore, the normalization operations obviously allow to reduce the dimensionality. Table 4.1 illustrates an example of text preprocessing pipeline and shows its effects on an Italian text snippet.

Documents filtering by concept

This step consists of applying a filter to select only documents containing a certain concept (e.g., through regex patterns). For instance, in this way it is possible to carry out the analysis taking into account only the documents mentioning a specific keyword (such as the name of a medical treatment

Step #	Transformation	Transformation Description	Text
0	/	/	Io soffto di acalasia e la prox sett avrò l’interventooo hellerDor :(https://bit.ly/2Tzdjcm
1	Encoding uniformization	Standardize encoding for all documents and handle accented characters	Io soffto di acalasia e la prox sett avro l’interventooo hellerDor :(https://bit.ly/2Tzdjcm
2	Symbols normalization	Standardize symbols for all documents	Io soffto di acalasia e la prox sett avro l'interventooo hellerDor :(https://bit.ly/2Tzdjcm
3	Emotes normalization	Replace emotes with standard sentiment tags	Io soffto di acalasia e la prox sett avro l'interventooo hellerDor EMOTEBAD https://bit.ly/2Tzdjcm
4	Attached words splitting	/	Io soffto di acalasia e la prox sett avro l'interventooo heller Dor EMOTEBAD https://bit.ly/2Tzdjcm
5	URLs removal	/	Io soffto di acalasia e la prox sett avro l'interventooo heller Dor EMOTEBAD
6	Case-folding	Convert text to lower case	io soffto di acalasia e la prox sett avro l'interventooo heller dor emotebad
7	Internet slang translation	Convert abbreviated internet slang into the corresponding term or phrase	io soffto di acalasia e la prossima settimana avro l'interventooo heller dor emotebad
8	Extra spaces removal	Remove extra spaces at the beginning, at the end and between words	io soffto di acalasia e la prossima settimana avro l'interventooo heller dor emotebad
9	Word lengthening fixing	Rip offs repeated characters more than two	io soffto di acalasia e la prossima settimana avro l'interventoo heller dor emotebad
10	Spelling correction	/	io soffro di acalasia e la prossima settimana avro l'intervento heller dor emotebad
11	Language translation	/	i suffer from achalasia and next week i will have the heller dor operation emotebad

Table 4.1: Example of quality preprocessing pipeline for Italian text

or a food). Despite its simplicity, this operation allows the possibility of distinguishing global analyzes (aimed at investigating the phenomenon in its entirety, regex “*”) from local analyzes (focused on a particular aspect).

Entities preprocessing

Similarly to the quality preprocessing phase on the textual component of the documents, it is necessary to carry out transformations also on the labels returned by the NER. These operations include encoding setting, case-folding and replacement of spaces with underscores (to avoid the separation of terms describing the same concept during tokenization).

Entities reconciliation

If the NER used takes into account the context of the terms for their recognition and labeling, a reconciliation phase may be needed. In fact, the NER system could label the same entity in different documents with dissimilar types. More specifically, if the Named Entity Recognizer adopts a hierarchical classification by assigning one or more types of a single branch to an entity, the same entity could be labeled with more or less specific types depending on the context (e.g., "Rome" in some documents could be labeled as "Capital" and in others only as "City"). Similarly, if the Named Entity Recognizer is able to assign types of different branches to an entity, the same entity could be labeled with more or less types depending on the context (e.g., "Rome" in some documents could be labeled as "City;Sport" and in others only as "City").

Despite being very useful, a NER system can therefore create information fragmentation (albeit correctly in some cases). The same concepts labeled differently are seen as different concepts during the analysis. Labeling produced by a NER, if not managed, can have the effect of significantly increasing the terms to be considered in the process (as if the documents used different words from each other). The goal of the reconciliation phase is to reduce heterogeneity for terms related to the same concept (i.e., reduce the number of concepts with different tags). This offers the possibility of not limiting the power of the language model, highlighting latent correlations, and consequently improving the results obtainable from the descriptive analysis.

The resolution of different types within the same hierarchical branch can be done with generalization and specialization operations. Depending on the macro-type (e.g., place, food), it may be more appropriate to standardize the tags with the more specific one identified on all documents or with the more general one. For instance, for places it might be appropriate to bring all types to the lowest hierarchical level (if "Rome" is labeled once as "Capital" and the others as "City", all labels are replaced with "Capital"). Since a type-node may have multiple children, the specialization operation is more complex than that of generalization and requires context. A simpler approach (which does not take semantics into account), even in the case of mixed types among multiple hierarchical branches, is to choose the most frequent label on all documents.

Entity tagging in documents

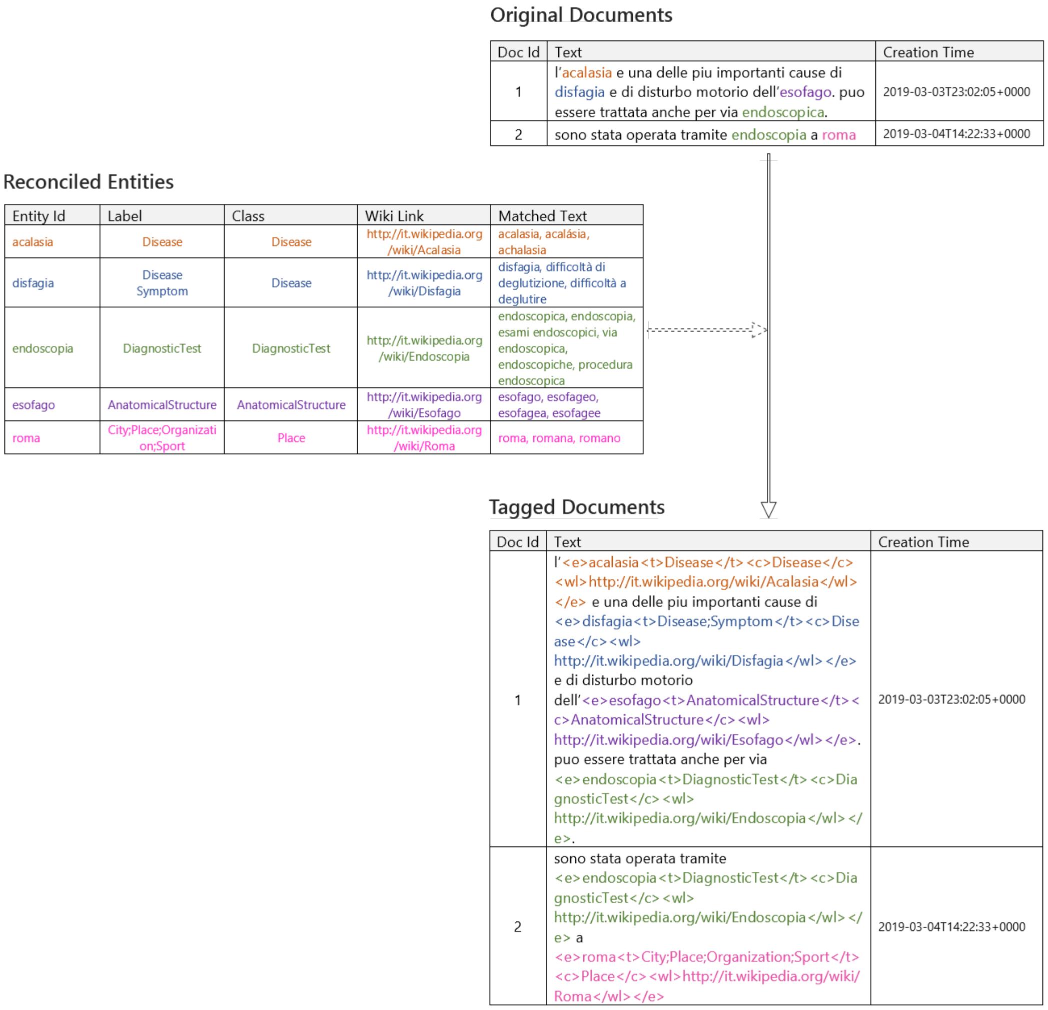
After having carried out the reconciliation phase of the entities recognized by the NER, the standardized information of the entities themselves must be reported directly in the textual content of the documents. To this end, the text that allowed the NER system to identify a certain entity (also known as "matched text") is replaced with a tag containing all the necessary data (e.g., entity id, entity types, ids for external knowledge bases). If the NER can assign more than one type to an entity, it may be useful (especially for visualization purposes) to insert a summary macro type inside the tag.

It is important to note that a NER system may not always be able to identify a concept in all the documents in which it appears (e.g., "Rome" could be recognized in certain documents but not in others if the context is not sufficient to allow it). Together with the problem of multiple labeling for the same concept (already addressed with the reconciliation phase), this fact contributes significantly to the reduction of the language model's power (with a much lower number of labels than the actual ones). In order to solve this weakness in an automated way, all the terms making perfect lexical matching with one of the matched texts recognized by the NER system for a certain entity are also labeled. In this way, if the Named Entity Recognizer detects "Rome" as a place entity in a certain document, all the terms "Rome" in the documents are labeled in the same way (also if not originally recognized). Although this correction applied to the results of the NER is likely to make some mistakes, overall the resulting benefits should be significantly more. An example of this stage with documents in Italian language is shown in Figure 4.4.

Lemmatization

Lemmatization allows to increase the similarity between the terms and therefore their frequencies. For most applications, where it is important to keep the meaning of the word, lemmatization is preferred over stemming. For example, "meeting" (noun) and "meeting" (verb) would be both stemmed to "meet", thus losing its original meaning, while the respective lemmas would be "meeting" and "meet".

Within the method, lemmatization is not performed before the application of the NER, in order not to compromise its effectiveness. In fact, as already mentioned, it is directly the NER system to include this operation if it is considered necessary according to its approach. Similarly, lemmatization is positioned after quality preprocessing to benefit from the corrections made on text. Obviously, for the terms representing NER tags, lemmatization is applied only to the entity identifier (i.e. its standardized name).



Original Documents

Doc Id	Text	Creation Time
1	l'acalasia e una delle piu importanti cause di disfagia e di disturbo motorio dell'esofago. puo essere trattata anche per via endoscopica.	2019-03-03T23:02:05+0000
2	sono stata operata tramite endoscopia a roma	2019-03-04T14:22:33+0000

Reconciled Entities

Entity Id	Label	Class	Wiki Link	Matched Text
acalasia	Disease	Disease	http://it.wikipedia.org/wiki/Acalasia	acalasia, acalásia, achalasia
disfagia	Disease Symptom	Disease	http://it.wikipedia.org/wiki/Disfagia	disfagia, difficoltà di deglutizione, difficoltà a deglutire
endoscopia	DiagnosticTest	DiagnosticTest	http://it.wikipedia.org/wiki/Endoscopia	endoscopica, endoscopia, esami endoscopici, via endoscopica, endoscopiche, procedura endoscopica
esofago	AnatomicalStructure	AnatomicalStructure	http://it.wikipedia.org/wiki/Esofago	esofago, esofageo, esofagea, esofagee
roma	City;Place;Organization;Sport	Place	http://it.wikipedia.org/wiki/Roma	roma, romana, romano

Tagged Documents

Doc Id	Text	Creation Time
1	l'<e>acalasia<t>Disease</t><c>Disease</c><wl>http://it.wikipedia.org/wiki/Acalasia</wl></e> e una delle piu importanti cause di <e>disfagia<t>Disease;Symptom</t><c>Disease</c><wl>http://it.wikipedia.org/wiki/Disfagia</wl></e> e di disturbo motorio dell'<e>esofago<t>AnatomicalStructure</t><c>AnatomicalStructure</c><wl>http://it.wikipedia.org/wiki/Esofago</wl></e>. puo essere trattata anche per via <e>endoscopia<t>DiagnosticTest</t><c>DiagnosticTest</c><wl>http://it.wikipedia.org/wiki/Endoscopia</wl></e>.	2019-03-03T23:02:05+0000
2	sono stata operata tramite <e>endoscopia<t>DiagnosticTest</t><c>DiagnosticTest</c><wl>http://it.wikipedia.org/wiki/Endoscopia</wl></e> a <e>roma<t>City;Place;Organization;Sport</t><c>Place</c><wl>http://it.wikipedia.org/wiki/Roma</wl></e>	2019-03-04T14:22:33+0000

Figure 4.4: Labeling of documents with reconciled entities (Italian text)

Documents classification

The way in which a certain class is distributed on documents represents a phenomenon that can be investigated with a descriptive analysis, to understand the reasons behind it. From this point of view, the attribute to be considered as a class could already be available within the dataset, or it could be calculated for each document at this stage. In any case, even if the classification technique should be supervised, the descriptive text mining method presented in this document is not.

This flexibility means that applications can be truly numerous. The classification could refer to a category of air accidents and the description could be aimed at investigating the factors that lead to destructive ones. Similarly, the classification could be understood as an opinion mining task on restaurant reviews, and the description in this case could have the objective of understanding why the opinion of a certain restaurant is overall negative. Furthermore, the classes could concern politicians, and the descriptions could be aimed at capturing the most *semantically* representative terms of their speeches (without the limits of the simpler approaches, based only on counting of occurrences or term weightings).

Analysis preprocessing

Unlike quality preprocessing, analysis preprocessing has the objective of preparing data for analysis. It typically includes transformations such as case-folding, replacement of punctuation and numbers with spaces (except for entity tags), extra white-spaces removal and stopwords removal. Depending on the application, the operations can be different and also concern the elimination of HTML tags, the conversion of number words into numeric form, and so on. The text is therefore normalized after labeling, reducing any form of noise (i.e., elements irrelevant for the purposes of the subsequent phases). An example is reported in Table 4.2.

Tokenization is another central aspect in this step. The descriptive text mining method illustrated in this document can be used in general with any N-Gram tokenization (e.g., Unigram, Bigram, Trigram), both at the word and character level. Although less intuitive, subword models have recently shown that they are capable of capturing and exploiting semantic text properties, and the results are very promising [178, 179, 180]. Even in the case of a statistical approach, character-level models can be very powerful, significantly increasing the correlations between tokens and therefore the ability to bring out latent associations. However, with N-Gram at character-level, it is necessary to face the problem of having to reconstruct the words (in order to return a meaningful description of the phenomenon). This issue can be solved with solutions based on matrix products. Since character-models would require additional steps, the next stages of the method refer to a unigram word-level approach for simplicity (remembering how it is still possible to adopt different tokenizations). With a unigram word-level model, each entity term also includes the tag with the reconciled data (obtained from NER results).

Term-document matrix construction

A term-document matrix is extracted from the corpus, where each row stands for a unique term t, each column stands for a unique document d, and each cell contains the frequency with which t appears in d. Consequently, the resulting occurrence matrix is large and typically very sparse. Along

Text after quality preprocessing and entities tagging	l’<e>acalasia<t>Disease</t><c>Disease</c><wl>http://it.wikipedia.org/wiki/Acalasia</wl></e> e una delle piu importanti cause di <e>disfagia<t>Disease;Symptom</t><c>Disease</c><wl>http://it.wikipedia.org/wiki/Disfagia</wl></e> e di disturbo motorio dell’<e>esofago<t>AnatomicalStructure</t><c>AnatomicalStructure</c><wl>http://it.wikipedia.org/wiki/Esofago</wl></e>. puo essere trattata anche per via <e>endoscopia<t>DiagnosticTest</t><c>DiagnosticTest</c><wl>http://it.wikipedia.org/wiki/Endoscopia</wl></e>.
Text after lemmatization	il <e>acalasia<t>Disease</t><c>Disease</c><wl>http://it.wikipedia.org/wiki/Acalasia</wl></e> essere una del piu importante causa di <e>disfagia<t>Disease;Symptom</t><c>Disease</c><wl>http://it.wikipedia.org/wiki/Disfagia</wl></e> e di disturbo motorio del <e>esofago<t>AnatomicalStructure</t><c>AnatomicalStructure</c><wl>http://it.wikipedia.org/wiki/Esofago</wl></e> puo essere trattare anche per via <e>endoscopia<t>DiagnosticTest</t><c>DiagnosticTest</c><wl>http://it.wikipedia.org/wiki/Endoscopia</wl></e>.
Text after analysis preprocessing	<e>acalasia<t>Disease</t><c>Disease</c><wl>http://it.wikipedia.org/wiki/Acalasia</wl></e> importante causa <e>disfagia<t>Disease;Symptom</t><c>Disease</c><wl>http://it.wikipedia.org/wiki/Disfagia</wl></e> disturbo motorio <e>esofago<t>AnatomicalStructure</t><c>AnatomicalStructure</c><wl>http://it.wikipedia.org/wiki/Esofago</wl></e> trattare <e>endoscopia<t>DiagnosticTest</t><c>DiagnosticTest</c><wl>http://it.wikipedia.org/wiki/Endoscopia</wl></e>

Table 4.2: Example of analysis preprocessing for Italian text, before tokenization

its rows there are both labeled (“entity”) and unlabeled (“standard”) terms. Maintaining standard terms as well as entity terms is important for describing a phenomenon. In standard terms, in fact, we can also find adjectives (e.g., “good”, “bad”) and highly specific terms of a domain (not recognizable by a non-specialized NER system). These types of terms are obviously fundamental for interpretation purposes. Having labeled versions for terms corresponding to entities constitutes an added value as regards method effectiveness, quality of the result, non-flat description, analytical possibilities and integration with other knowledge bases.

Feature selection

This stage involves the removal of sparse standard and entity terms, selecting only those of interest. Term selection is used for several reasons: simplifies the model and makes it easier to interpret, reduces training times, avoids the curse of dimensionality, and promotes generalization. Therefore, irrelevant or partially relevant terms can have a negative impact on the model. In this context, in addition to stopwords, terms with a significantly reduced number of occurrences (below a certain percentage threshold, such as 1%) can also be considered irrelevant and discarded. From a co-occurrence point of view, each term should appear at least twice. However, the percentage threshold for standard terms should be distinguished from that for entity terms. Depending on the domain and the specific data source, a concept of an interesting type (e.g., a drug) could be of fundamental importance for the purposes of the analysis even if it is scarcely mentioned in the documents. So, for example, the minimum presence threshold for entity terms may be lower than that for standard terms. In any case, this phase must be performed with caution because the more terms are eliminated, the more the latent correlations become weak.

Term weighting

Raw counts do not consider the significance a term has in the document in which it appears. To better represent the importance of each term in each document, term weighting methods are applied to the cell values of the term-document matrix (e.g., tf-idf). A good comparison of the available schemes is proposed in [181]. This is a crucial phase within the method as it strongly affects the production of the description for the phenomenon investigated. More specifically, it determines the norms of the vectors in the new space and therefore their impact on the result.

Word clouds and word frequency analysis

Word Clouds (or Tag Clouds) are visual representations of text data in the form of tags, which are typically single words (or terms) whose importance is shown by way of their size and color. The role of this technique (and that of word frequency analyzes in general) is secondary to that of other phases belonging to the proposed method. However, they can be useful to provide a general and quick overview of the data that can act as a guide for subsequent deep qualitative investigations. This acquires even more value considering the possibility (thanks to filters on concepts) to carry out more "local" analyzes. Moreover, a series of word clouds can be a good way to show changes in popular themes over time (e.g., terms used by patients when talking about a specific medical treatment, terms used in political speeches, terms used to compose newspaper headlines). This possibility of representing trends that are not exclusively "large" further justifies the importance of considering time-related filters. For what concerns frequency analyzes, it is known how these can be used to establish the "signature" of a certain author, the cultural level of the writer, its use of technical jargon or slang, and other writing features. Obviously, word clouds also have different limits and are not suitable for performing accurate analysis. It should be noted that, despite their traditional use consider only the frequency of a word to judge its importance (without managing context, meaning, and derivations), in general they can also take into account any weighting schemes applied. For this reason they are placed at this point in the method. So, they are ideal for exploring large amounts of text and creating simple informative visualizations, shareable and easy to understand. Within a descriptive analysis, the word clouds of documents belonging to different classes can also be compared with each other. They can also be used to show the most frequent terms for each macro typology recognized by the NER system. Although based only on the weighted occurrences of terms, these tools can offer a first idea of the most significant terms in both classes.

Language Model and Latent Semantic Mapping

In this phase the method involves the application of a language model (LM), which forms the basis for the whole analysis. LMs go beyond the representation of the text in a machine-understandable form. They are techniques of transformation and dimensionality reduction that allow the construction of a *semantic vector space* from which perform knowledge extraction, such as the detection of latent similarities between terms and documents.

LMs can be constructed in three ways:

- algebraically (e.g., LSA + SVD);
- probabilistically (e.g., P-LSA, LDA);
- neural network based (word embeddings such as word2vec and BERT).

The terms and documents are thus embedded into a new common space with a *Latent Semantic Mapping* operation, depending on the chosen LM. For example, with LSA the positions of all terms and documents in the new space are obtained respectively from the products of the $U\Sigma$ and $V\Sigma$ matrices (resulting from SVD). In the case of a neural network-based LM, the concepts of word embeddings and document embeddings must be merged. There are several ways to calculate document embeddings. One of the most intuitive approaches to calculating a document embedding is to derive it directly from the execution of some arithmetic operations on all the vectors corresponding to the words of the document itself, in order to summarize them into a single vector in the same embedding space. Two such common pooling operators are average and sum.

Within the new space, the prevalent way to estimate the semantic similarity of two vectors (be these terms or documents) is to apply the *cosine similarity* between them (a measure commonly used in IR and Text Mining).

$$sim(A, B) = cos(\theta) = \frac{A \cdot B}{\|A\| \, \|B\|} = \frac{\sum\limits_{i=1}^{n} A_i B_i}{\sqrt{\sum\limits_{i=1}^{n} A_i^2} \sqrt{\sum\limits_{i=1}^{n} B_i^2}} \tag{4.1}$$

2D space representation

The visualization of terms and documents in the latent semantic space (built in the previous step) is useful for several reasons. Starting from it, it is possible to:

- identify the correlations between terms and terms, documents and documents, and terms and documents;
- have a graphical feedback on the distribution of documents based on their class;
- recognize the presence of any clusters;
- understand the effectiveness of the model and whether it is necessary to intervene again on the previous preprocessing phases to make adjustments.

Although the new space is made up of k dimensions, a 2D representation must be adopted to make the graph suitable for human observation. Under this point of view, *t-SNE* [182] can be useful for mapping high-dimensional data to two dimensions (compressing all the original ones so as to minimize divergences). However, this task can be achieved independently of it. For example, with LSA as LM, the choice of the two dimensions to be adopted for visualization purposes can be made directly from the singular values in the matrix Σ. Since the latter are in descending order and indicate the importance of their dimensions in the transformed space, a good choice concerns the dimensions associated with two high singular values without too much difference between them (to have a good approximation and avoid a strong crushing of data on one axis with respect to the other). In any case, generally it does not go beyond the fourth dimension, because the information captured is much lower than the previous ones. The

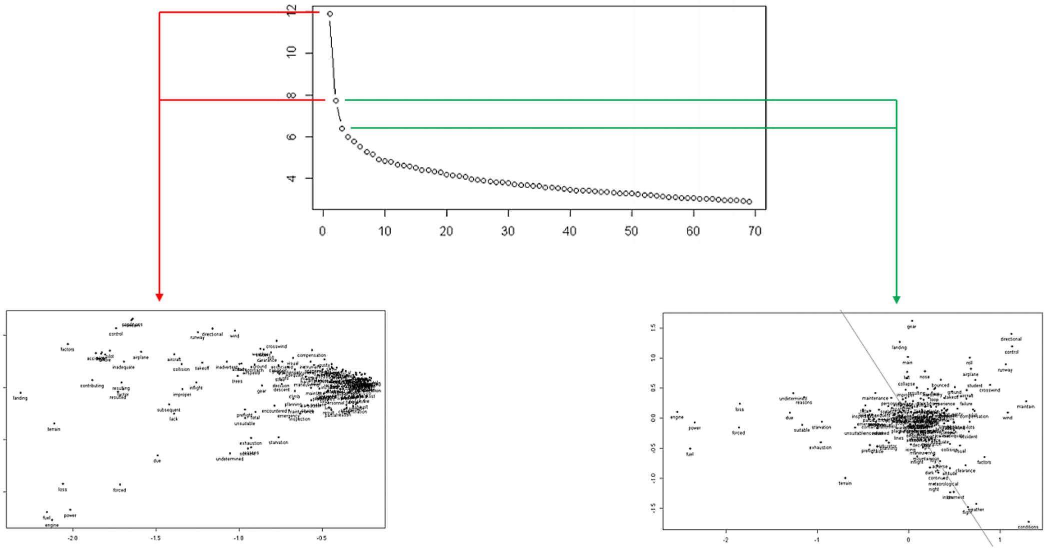


Figure 4.5: 2D visualization of the latent semantic space starting from the power law curve obtained after the SVD decomposition (LSA). In the example, the left image shows the distribution of the terms in the semantic latent space considering the dimensions 1 and 2, while the right one refers to the dimensions 2 and 3. In the first case there is very little variability on the second eigenvalue, and the resulting distribution is not satisfactory (takes the form of an ellipsoid). In the second case, the distribution is less concentrated and consequently it is better for visual observations.

power law curve formed by the singular values in Σ is a valid tool to make this decision, as shown in Figure 4.5.

The coordinate values of the terms may have a wider range than that of the documents. To prevent terms and documents from being displayed at different scales, a good practice is to *normalize* vectors (make their length 1).

From the graph it is possible to identify the semantically correlated terms on the basis of their cosine similarity (Figure 4.6). The more a pair of terms forms a small angle with the origin, the greater the similarity between them (i.e., terms placed along the same straight line). These terms often appear together in documents because they are synonyms or often associated. However, having compressed an originally high-dimensional space into only two dimensions, close terms in the graph could actually be far away considering also the other unselected dimensions (which is why the use of t-SNE is recommended).

In order to better understand the distribution of terms in space and the quality of the correlations, at the time of visualization the entity terms can be colored according to their macro-class or their detailed combination of types

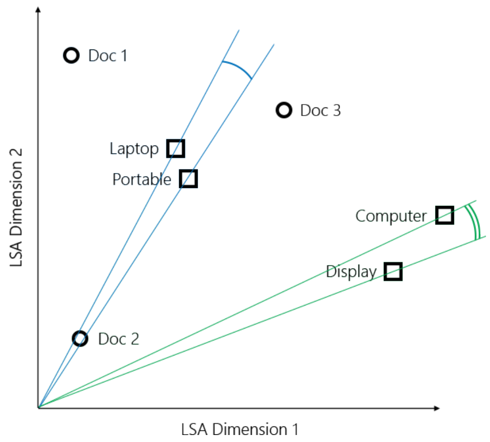


Figure 4.6: Similarity between terms in the latent semantic space

(obtained as result of the reconciliation phase). Similarly, documents can also be colored according to their class. A complete graph of the latent space can therefore be obtained by considering the overlapping representation of normalized terms and documents, accompanied by their class (Figure 4.7).

Selection of the number of dimensions

If the use of only two dimensions is suitable for visualization purposes, in the rest of the analysis it could involve a significant loss of information. In fact, due to the strong approximation of the data, the distinction between terms that are not semantically related could be lost.

In general it is necessary to choose an optimal number k of dimensions to be used in the analysis. If k is too high, many details are retained and noise in the data does not reveal semantic correlations between terms. If k is too low, those dimensions that differentiate distinct terms are also removed, causing incorrect correlations to emerge.

The three classes of LMs have different ways to do it. In SVD you choose the rank. In word embeddings, the size of the vectors is defined [183]. In P-LSA there are probabilistic parameters with which it is possible to reduce the number of dimensions. In LDA you choose the number of topics and the number of keywords for each topic.

Specifically, for what concerns LSA with SVD, the choice refers to the k dimensions with the highest singular values, as they represent the most recurrent correlations in the analyzed data. So, after applying the decomposition by choosing a number of dimensions for the new space (through heuristics, reasoning about the domain, or other approaches), a further reduction in dimensionality can be made to lighten the computational load required by the

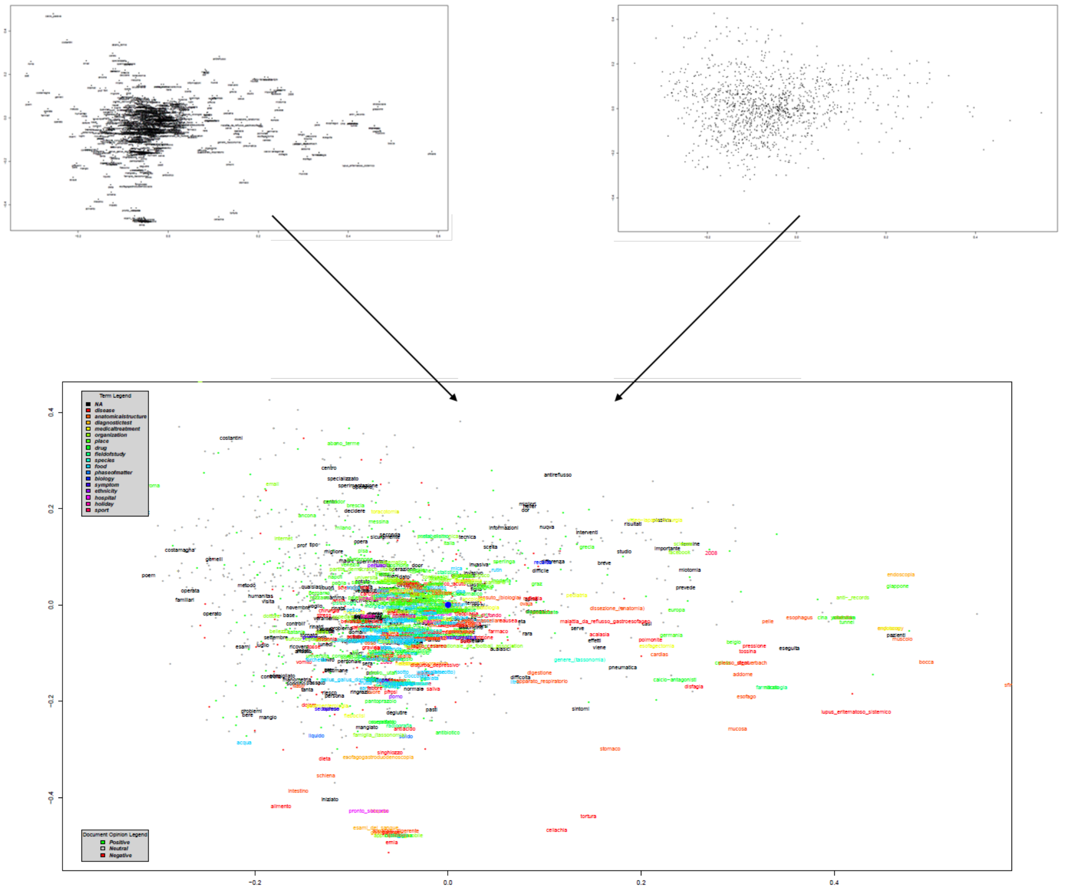


Figure 4.7: Graph of the latent semantic space with normalized and overlapping terms and documents

subsequent analytical phases. To decide an optimal value of the number of dimensions to use, it is possible to analyze the descending sequence of singular values to search for a *knee point* in the progression. This can be done both visually and formally. In fact, a knee point is a point where the radius of the *curvature* of the function that interpolates the hyperbola corresponds to a local minimum. Considering that the curvature of a function $y = f(x)$ is $c = y''/(1 + (y')^2)^{3/2}$, a valid number of dimensions k can therefore coincide with one of the local minima (Figure 4.8). The idea is that the informative contribution given by the dimensions associated with the eigenvalues that follow a knee point is lower, making an approximation possible.

In general it is advisable to perform tests with multiple ranks by selecting some of the potentially optimal ones, with this or other methods. In any case, in practice it has been seen that it is not necessary to use a number of eigenvalues greater than 300. One way to conduct these tests is to verify the consistency

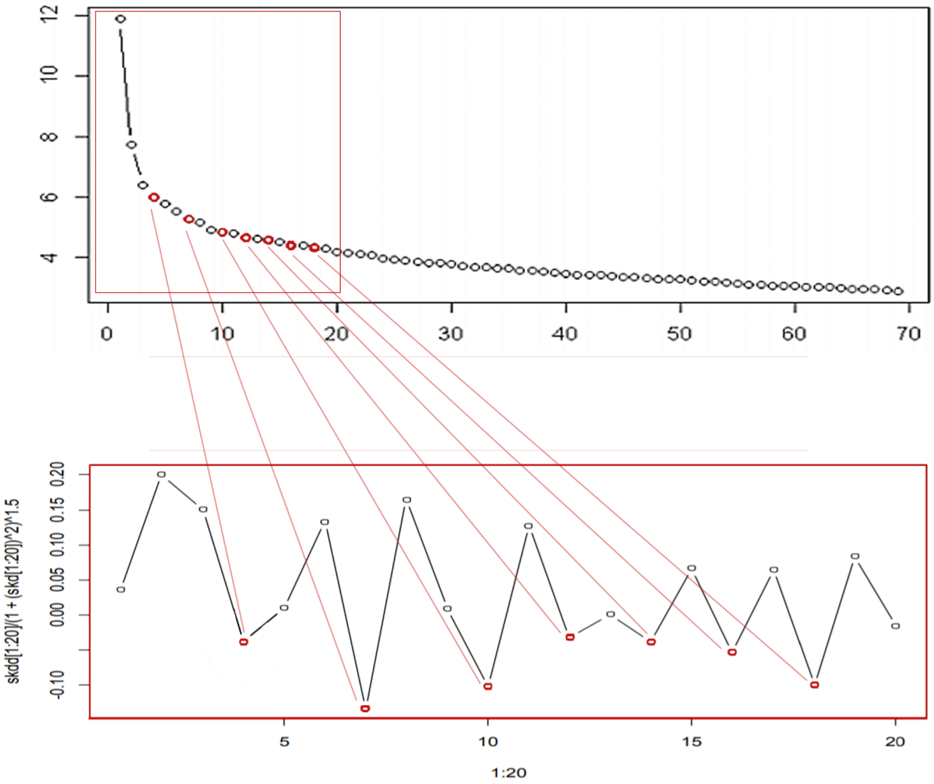


Figure 4.8: Selection of the number of dimensions in LSA, through the search for knee points in the curve of the sequence of singular values

with respect to a particular query of the first N documents semantically most similar to it, and so the precision of the query itself (this step will be explained later). For example, given a query mentioning a place, the documents most similar to it are expected to have textual content related to this place. Since SVD (and so LSA) is based on co-occurrences, it is important to note that during these checks, real positives documents (actually related to the query) may be listed although these do not directly contain the query term(s). This could not happen with a Boolean research model where one works lexically and not semantically.

The operations described in the following phases involve only the dimensions selected in this step.

Interesting correlations

Once the correct number of dimensions has been ascertained, some interesting correlations can be quickly searched (e.g., the terms / documents most similar to the concept of interest in the case of local analysis).

Qualitative analysis of the graph to identify unexpected concentrations

In the case of uniform distribution, imagining to divide the space into quadrants, a quantity of documents proportional to the original one should be found for each class. The areas of the space outlined by the LM in which there are unexpected concentrations (different from those foreseen in the case of random distribution), indicate the presence of elements of interest for the analysis (Figure 4.9). By researching which terms are found in these areas, it is possible to interpret them to identify the causes that contribute to the phenomenon. This phase of the method therefore has the aim of recognizing the possible presence of areas to be investigated.

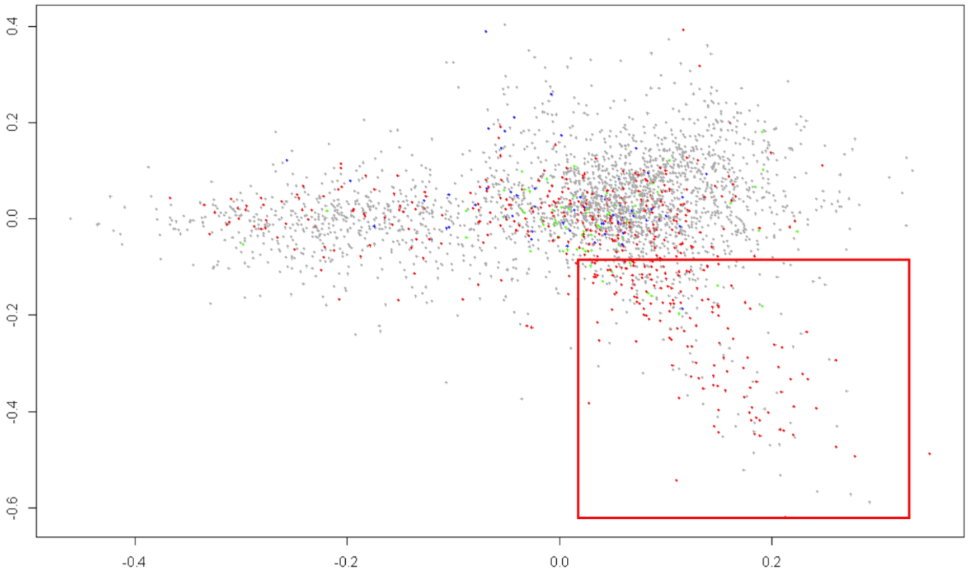


Figure 4.9: Example of unusual concentration of documents belonging to a certain class, in the 2D representation of the latent semantic space

Iterative search of terms related to documents of a certain class

This last phase represents the heart of the descriptive analysis method proposed in this thesis. By following the operations explained below, it is in fact possible to construct a *probabilistic description* (step by step) for the phenomenon we have chosen to analyze. In particular, the resulting description will consist of the set of terms that best characterize the distribution of the class representing the phenomenon in the latent semantic space.

First, the most representative term must be visually identified in the area highlighted in the previous stage. In doing this, it is necessary to focus on the terms placed in a central position within the area itself and at a greater distance from the origin (with a high norm and therefore a high relevance). Figure 4.10 shows an example.

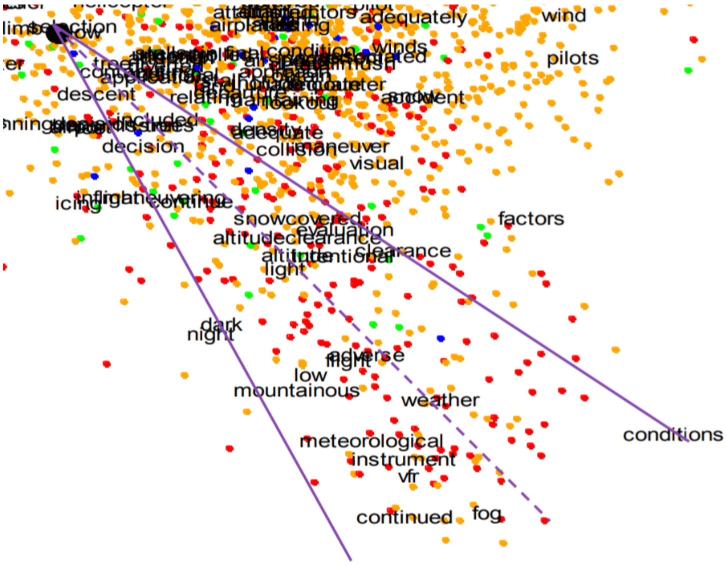


Figure 4.10: Example of first-term selection, according to the proposed descriptive text mining method. As can be seen, "weather" is among the terms with the highest norm, positioned at the center of the concentration of documents labeled with the class associated with the red color. Consequently, it is a valid choice.

The selected term (and in general the set of key terms to be composed) can be seen as a query. In fact, a query is made up of text and can therefore be equated with a document, which can be folded-in[2] to the transformed space. For what concerns LSA, as said in 2.1.3, the representation of a query in vector form is dependent on that of the original documents: the terms considered in the vector must be the same and must be weighted with the same criterion. The query is first transformed into a vector indicating the presence/absence of each term (bag of words), and then subjected to the same term weighting scheme used initially. The query vector q can be converted to a new line of the matrix V with the formula $q_k = q^T U_k \Sigma_k^{-1}$. Instead, the position of the query in the latent space is obtained with $q^T U \Sigma^{-1} \Sigma = q^T U$ (similarly to the product $V\Sigma$ for mapping documents in the transformed space). The most relevant documents for a query are the vectors of the $V\Sigma$ matrix having a high cosine similarity with the query itself.

[2] Folding is the process of adding new documents to a space after its construction.

In order to mathematically demonstrate the correlation between the query q (currently made up only of the arbitrarily chosen term) and the class c representing the phenomenon for which an explanation is sought, the *chi-squared (χ^2) test*[3] can be used. In particular, we do not want to verify the lexical correlation in the original space (where only the documents mentioning t would be considered), but the semantic one in the transformed space. Consequently, we do not adopt a Boolean search model, but a *ranking search model* (in which the results are sorted on the basis of a coefficient). To this end, it is necessary to set the maximum number of ranked results (i.e., semantically more similar documents to the query) to be considered (*R-precision*). In establishing the threshold we must give the query the opportunity to retrieve all the documents it wants to represent. Consequently, R-precision is valued with the number of documents in the class linked to the phenomenon being investigated.

So, in this case the two events correspond to the occurrences of documents satisfying q and c, within the R-precision. Starting from the contingency tables (Table 4.3) with the observed and estimated frequencies for the combinations of the two events and their complementaries, the correlation can be calculated in the following way.

	$e_c = 0$	$e_c = 1$
$e_q = 0$	N_{00}	N_{10}
$e_q = 1$	N_{01}	N_{11}

	$e_c = 0$	$e_c = 1$
$e_q = 0$	E_{00}	E_{10}
$e_q = 1$	E_{01}	E_{11}

Table 4.3: Contingency tables structure for chi-squared test between a query and a class

$$\chi^2(\mathbb{D}, q, c) = \sum_{e_q \in \{0,1\}} \sum_{e_c \in \{0,1\}} \frac{(N_{e_q e_c} - E_{e_q e_c})^2}{E_{e_q e_c}} \tag{4.2}$$

where:

$\mathbb{D}$ = corpus (repository of documents)
q = query
c = class
e_q = variable indicating the presence or absence in the top-R identified by q
e_c = variable indicating the presence or absence of the class in the corpus
$N_{e_q e_c}$ = number of documents observed with e_t and e_c
$E_{e_q e_c}$ = number of documents expected with e_t and e_c

[3] A general and common statistical method for verifying independence between two events under the null hypothesis, providing reliable answers for researchers [184].

In other words, the formula estimates the difference between the observed and expected values for q and c in $\mathbb{D}$, assuming that q and c are independent. While the number of documents observed is directly reflected in the data, the expected frequencies are calculated as $E_{e_q e_c} = |\mathbb{D}| \cdot P(t) \cdot P(c)$. For example, $E_{11} = |\mathbb{D}| \cdot ((N_{11} + N_{10})/|\mathbb{D}|) \cdot ((N_{11} + N_{01})/|\mathbb{D}|)$.
The higher χ^2, the lower the probability that the hypothesis of independence between t and c holds. To establish whether an index value is high or low (i.e. whether the null hypothesis is rejected or not), a *threshold* must be set. To do this, the degrees of freedom must be considered, calculated as $df = (i-1)(c-1)$. For each degree of freedom there is a *distribution table* of χ^2 which indicates the confidence associated with the index value. The one for degree of freedom equal to 1 (as in the method described here) is shown in Table 4.4. So, if the χ^2 between the chosen term and the class associated with the unusual concentration is greater than a critical value (e.g., $\chi^2 > 6.63$ and independence probability $< 1\%$), the null hypothesis can be rejected. In this case, let n_c be the number documents belonging to class c in top-R and $|c|$ the number of instances of class c (used as R-precision), it is possible to say that the query characterizes a number of instances related to the phenomenon being described equal to $n_c/|c| \cdot 100$, with a probability corresponding to the p-value for the calculated χ^2.

p-value	χ^2
0.1	2.71
0.05	3.84
0.01	6.63
0.005	7.88
0.001	10.83

Table 4.4: Distribution table χ^2 with 1 as degree of freedom

The effectiveness of the query can also be ascertained graphically, displaying the distribution of the classes for the top-R retrieved documents (i.e. sorted by decreasing semantic importance), as shown in Figure 4.11.

After having verified with a formal approach the relevance of the term chosen for the class linked to the phenomenon for which an explanation is sought, it is possible to proceed with the extension of the description (currently only consisting of this first term).

The analysis therefore continues with the search for terms closest to the query. For each term u_i in U, we calculate the similarity between u_i and q as $cos(u_i \Sigma_k^{1/2}, \Sigma_k^{-1/2} U_k^T q)$. Among the terms with higher similarity, we choose one that also has a high norm and we insert it as a second element in the query. Like before, the chi-squared test is rerun to verify the correlation between

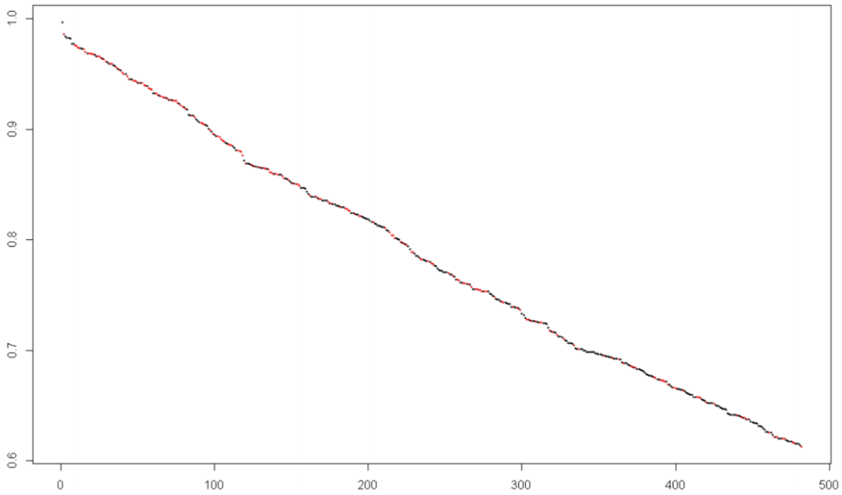


Figure 4.11: Example of R-precision visualization for a query. The red dots indicate the relevant documents, and their concentration in the first results confirms the correlation between the query and the class.

the new query and the class of interest. To verify that the query gave results similar to those of the previous step, we can see the number of documents in the intersection. The process is repeated iteratively, as long as the correlation between the query and the class does not fall below a set p-value threshold. The query is progressively made more specific, observing the terms most relevant to it and therefore most often associated with the phenomenon. Once the analysis is complete, the method allows you to obtain a set of terms similar to a sentence (e.g., “adverse weather instrument meteorological continued vfr” in a domain linked to plane crash reports). Figure 4.12 shows an example.

Finally, some observations on visualization approaches and Language Models - different from those adopted as examples - are reported.

- The interactive method guided by the observation of the 2D graph related to the latent semantic space cannot be directly replicated with a visualization realized through t-SNE. In fact, this algorithm tends to produce clusters and the distance metric commonly used with it is the Euclidean distance. However, it is an excellent solution to bring out similarities between the terms, to be used in conjunction with the proposed technique. Given its ability to take into account all dimensions, in fact, it can lead to different results and integration is of undoubted interest. In addition, it should be noted that, despite some still prototype attempts focused on position prediction, it is currently not possible to fold-in new documents

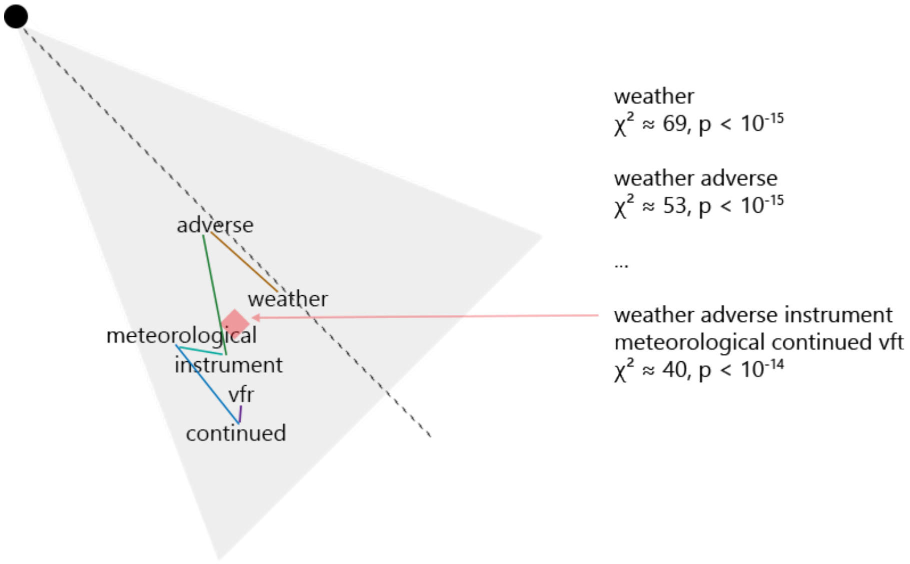

weather
$\chi^2 \approx 69$, p < 10^{-15}
weather adverse
$\chi^2 \approx 53$, p < 10^{-15}
...
weather adverse instrument meteorological continued vft
$\chi^2 \approx 40$, p < 10^{-14}
adverse
weather
meteorological
instrument
vfr
continued

Figure 4.12: Example of description construction. At each step the query is enriched with a semantically close term, and the correlation with the class tends to decrease.

on a t-SNE map without having to perform a new training.

- Fold-in is also possible with other Language Models, in addition to LSA. This property in any case is precious and not obvious. With Word Embeddings, in fact, it is possible to load a pre-trained model and use it without having to re-train it.

Evaluation

There are several ways to evaluate the correctness of the results produced by the method in its various parts. One could interpret the description itself and compare it with existing scientific reports or data. A more formal approach is based on the definition of a file containing a set of *gold standards* (i.e. known correlations expressed by an expert user). It can be made up of elements having a structure of this type: $x_1, \ldots, x_n \rightarrow y_1, \ldots, y_m$. For example, $smartphone, apple \leftrightarrow iphone$. Consequently, each correlation generally involves two documents (sets of terms). The effectiveness of the model can be proven by researching the semantic similarity between these documents in the latent semantic space and verifying that it is above a certain threshold. Again, clusters of similar terms in the transformed space can be manually evaluated. It is worth noting that, in the case of LSA, the use of a purely mathematical model makes the method entirely explainable. Observations proven by statistical tests are directly reflected by the co-occurrences of the data (regardless of whether the semantic similarities highlighted are also reflected in other sources).

4.3.2 Automatic knowledge extraction

This section describes a possible solution to fully automate the execution of the presented method.

Dimensionality calibration

A fundamental point within the analysis is the choice of the number of dimensions to work with. This concerns both the transformed space generated by the language model and the reduction of dimensionality aimed at lightening the computational load. Making this decision according to the domain is not always easy, and for automation purposes it cannot be standardized for all application cases.

Within this thesis we argue that an interesting approach is that focused on the reuse of gold standards as training set. The idea is to iterate over the parameters, make multiple attempts and choose the values that most bring out the known correlations. In particular, some known facts can be used for *calibration*, while the rest as *tests*. In addition, they can be used in *k-fold cross validation*. This mechanism simulates that of back-propagation (which would be done automatically using deep learning methods). Similarly to the evaluation phase, it can be carried out by calculating the semantic similarity between the pairs of documents involved in the correlations, for the various combinations of parameters. Compared to more basic implementations (such as the choice of the parameters returning the maximum average of the cosine similarities obtained on the permutations of the terms belonging to known correlations), this procedure is efficient, coherent with the analysis, and capable of overcoming the problem related to work with only pairs of terms. This semi-supervised approach (requiring a very small amount of information) allows the choice of the most suitable dimensions for the modeling of terms and domain documents, obviously also providing the opportunity to discover new correlations.

Clustering similar terms

Visual identification of term clusters directly from the 2D graph of the latent semantic space is difficult, imprecise and time consuming. However their extraction can be very useful in certain contexts. In LSA this corresponds to finding semantically similar groupings of row vectors in $U\Sigma$.

To automate this process a *Hierarchical Cluster Analysis (HCA)* is performed. HCA allows the discovery of relationships between data in an unsupervised way. The agglomerative strategy (bottom-up) puts each data point in its own cluster, identifies the closest two clusters and combine them into one, and repeats this step till all the data points are in a single cluster [185]. For the purpose of the descriptive text mining method, the cosine distance (orthogonal to the cosine similarity) is used as *metric*. Instead, the distance between sets of observations can be calculated with several *linkage criteria* (e.g., minimum, maximum, unweighted average, weighted average, centroid). From this point of view, a very common criterion in libraries is Ward [186]. Many implementations of HCA require to provide data in the form of *distance matrix*, where the number in the i-th row j-th column is the distance between the i-th and j-th terms. So, a distance matrix is symmetric and its diagonal elements are zero. *Heatmap* can be used to display it. Finally, the main output of this kind of analysis is a *dendrogram*, which shows the hierarchical relationships between the clusters. Figure 4.13 shows an example.

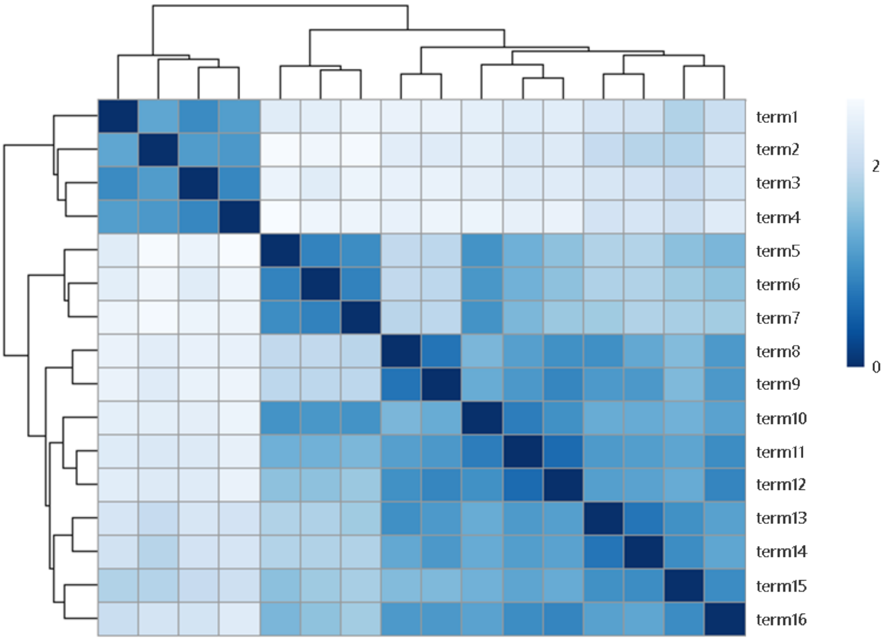


Figure 4.13: Example of heatmap and dendrogram obtained from a cosine distance matrix

Starting with the dendrogram, term clusters can be calculated in various ways. Some of these require the specification of a parameter. In the contexts in which the proposed method can be applied is unlikely that the number of clusters is known a priori. So, one possible way to automatically obtain these groupings is to cut the hierarchy to a certain height h. A more interesting approach is the one proposed by Suzuki et al. [187], focused on assessing the uncertainty of an HCA. For each cluster in HC, p-values are calculated via multiscale bootstrap resampling. Higher is the number of bootstrap replications, more confident is the estimation accuracy of p-values. So, the p-value of a cluster is a quantitative value between 0 and 1, which indicates how strong it is supported by data. In this case, the only parameter required is the minimum p-value of the term clusters to be searched (e.g., 0.95).

HCA can also be applied to automatically obtain the clusters of terms recognized by t-SNE, but with the Euclidean distance as a metric.

Relevant term adjacencies computation

By calculating the correlation between each pair of terms, it is possible to build a very useful data structure for subsequent tasks (e.g., query auto-completion) and researches. For each term, in fact, it is possible to keep track of the top N terms related to it, in descending order and with a positive correlation above a certain threshold. The inverse correlations between terms, in fact, are not always significant and, depending on the specific case, it is possible to evaluate whether to consider them or not.

The correlations can be expressed in the form of probability ($p \in [0, 1]$). Using P-LSA as LM, these data would be immediately available. Alternatively, there are two main ways to calculate them.

- Cosine similarity (remapped from $[-1, 1]$ to $[0, 1]$).
- P-value associated with chi-squared test.
  This represents a more formal and rigorous theoretical model from a mathematical point of view. It consists in calculating the χ^2 index starting from a 2x2 contingency matrix between two terms (instead of a term and a class). With this approach, we observe and estimate the number of times the two terms appear together, do not appear or appear alternately in the latent semantic space.

Starting query

Within a latent semantic space, there may be multiple areas of concentration (with non-random distributions) linked to the documents of the class of interest, and not necessarily just one.

When considering the automation of the method, it is therefore necessary to identify an approach that allows repeating the analysis on multiple areas (recognizing them through clustering) or that provides the user with the possibility of indicating the branch from which to start (through a query). The latter mode is very interesting. In a medical domain, for example, a patient can initially express some personal information (such as symptoms and treatment performed), and then see if there are terms semantically associated with them, based on the experience of other patients. Even if the description could not be further enriched without falling below the established p-value threshold, this still constitutes a result (absence of correlations). As a result, even queries located in areas of low interest are not a problem.

First term selection

If a user-specified starting query is not available, the automated version of the method must still be able to manage the choice of the initial term.

This task can be solved by considering all the terms with norm higher than a certain threshold, and applying the chi-squared test on them in order to calculate their correlation with the class. The term chosen is the one with minimum p-value and highest norm.

The risk linked to the automation of this phase is the choice of a term that is not particularly interesting or significant. In this regard, a valid preprocessing is fundamental.

Choice of the next term

Another aspect of the method that needs to be automated concerns the choice of the term with which to enrich the description at each step.

This can be done by searching for the N terms semantically closest to the current query, but in any case respecting a minimum cosine similarity threshold. Among them, the term used to extend the query is the one with the highest norm, capable of reaching a description with a sufficiently low p-value.

PDF report

Although the automation of the proposed descriptive text mining method allows the subsequent realization of numerous applications (such as the knowledge graph learning, described in Chapter 5), having access to the intermediate information is still very useful to understand how the method itself came to a certain result. The automatic creation of a PDF report at the end of the analysis can therefore constitute an added value.

4.3.3 Modularity

Another strength of the proposed method lies in its flexibility. As observable from Figure 4.14, it consists of several modules that can be replaced according to the specific case, without compromising the analysis in its entirety.

- *NER.*
  The method is not strictly related to NER and its use is optional. Depending on the domain, it is possible to adopt the technology that is considered more appropriate.

- *Quality preprocessing.*
  For example, in social contexts it may be necessary to manage slangs and emojis, correct errors, and translate the text. Other scenarios may have different needs.

- *Analysis preprocessing.*
  Similarly to the previous case, it must be possible to apply specific transformations (often based on the language). Furthermore, a different tokenization technique could be adopted.

- *Document classification.*
  Depending on the type of phenomenon to be described, the classification logics can be manifold. For example, you might want to explain the reason behind negative comments about a particular concept. In this case the classification phase consists in the application of a opinion mining algorithm. Furthermore, within this specific subcategory, you might want to compare the results obtainable by multiple techniques. The method, although requiring classes, is independent of the single classification logic.

- *Language Model.*
  As already noted, the choice of the Language Model is crucial. The use of one approach rather than another can always be seen as a module. This is a great advantage for the proposed method, which therefore also allows to compare algebraic models, with those probabilistic or based on neural networks. From this point of view, there is a perfect 1-to-1 matching. All have as their objective the vector modeling of terms and documents, albeit adopting completely different solutions.

- *Visualization.*
  The 2D visualization techniques of the latent semantic space can be different. For example, we have seen how in LSA it is possible to select 2 dimensions starting from the observation of singular values. The use of t-SNE in this case becomes another component to be put on and off, to show what happens.

The combination of these modules with the setting of the various parameters associated with them, allows to respond to the needs of many application cases.

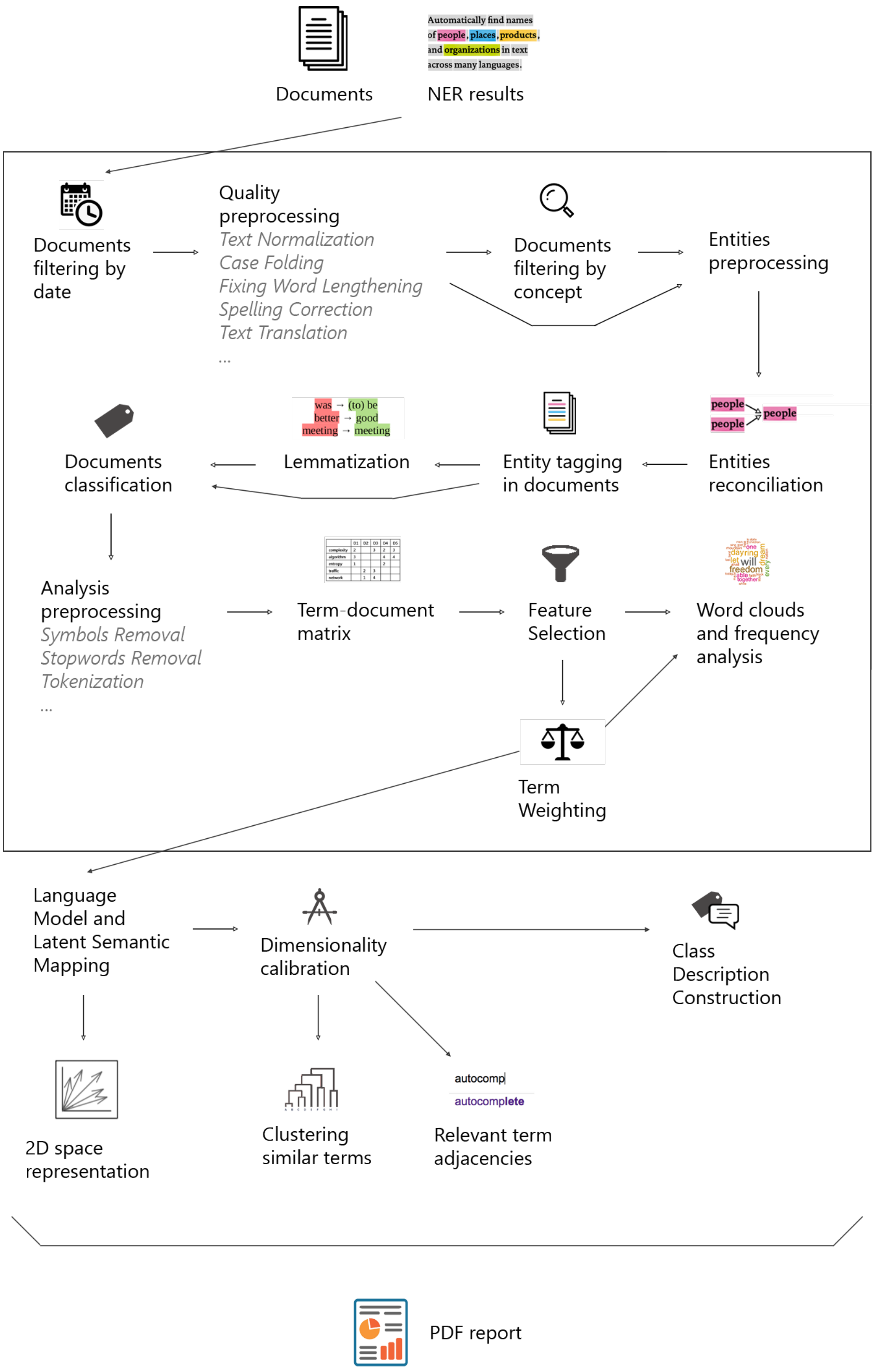


Figure 4.14: Proposed Descriptive Text Mining Method

Parameters summary

After providing a complete illustration of the text mining contribution, it is observed how the specific parameters depend in particular on the chosen LM. As for LSA, a summary of the inputs and the main parameters that can be considered for improving the results is reported in Table 4.5.

Parameter	Description
documents	The corpus of documents to analyze.
entities	Optional. The data of the entities extracted from the documents.
stopwords	Optional. The stopwords to remove from the dataset.
known_terms_correlations	The list of known correlated term groups.
qualityPreprocessingFun	The preprocessing function to be applied to documents in order to increase their quality.
concept	Regex for selecting only the documents mentioning a certain concept (local analysis). If set to "*", consider all documents (global analysis).
min_creation_date max_creation_date	Optional. Dates for carrying out the descriptive analysis only on documents belonging to a specific time window.
lemmatization	Indicates wheter the documents' texts should be lemmatized or not.
dimensionality_calibration_mode	The technique to use for dimensionality selection (e.g., maximum average of cosine similarity between all term permutations inside known correlations, or cosine similarity between folded-in documents).
starting_query	Optional. The document query from which start during descriptive analysis.
classificationFun	The function to apply for document classification.
min_standard_term_freq min_entity_term_freq	The minimum value of standard / entity feature frequency, expressed as percentage (e.g. 0.01 ->at least 1%).
min_term_norm_threshold	The minimum norm that must be possessed by a term to be considered relevant.
min_similarity_closeness	The minimum value for cosine similarity that must be satisfied during step-by-step analysis.
n_query_neighbors	The number of terms similar to the query to research.
min_pvalue_threshold	The minimum pvalue that must be satisfied by a class description.

Table 4.5: Main parameters for the descriptive text mining method based on LSA as LM

4.3.4 Final Considerations

By identifying the most representative terms of a phenomenon (intended as a distribution of a certain class, typically different from randomness), the proposed method provides a global explanation.

Moreover, it produces a description similar to a natural language sentence (an aspect particularly appreciated in literature).

Although the final result needs to be interpreted (requiring domain expertise), the approach is very powerful and flexible, opening the door to experimentation with different combinations of technologies and implementations. The joint use of LSA with modern solutions to different NLP tasks offers still unexplored possibilities.

The correlations between documents, terms and arbitrary queries can be analyzed numerically for confirmation.

The main part of the solution is totally unsupervised. Known correlations are required only if you want to automate the choice of the optimal number of dimensions to work with, adopting an original approach focused on semantics. In fact, the calibration phase is not essential and could for example be replaced with the use of heuristics and with the choice of the first minimum in the curvature. In any case, the amount of data required is minimal and inexpensive to produce compared to that frequently needed by other solutions.

Chapter 5

A New Unsupervised Methodology for Knowledge Graph Learning

"Our most pressing problems
require multidisciplinary solutions."
- Jim Breyer

In recent years, knowledge graphs have gained a lot of popularity for their ability to organize and make easily accessible the ever-increasing amount of digital information, enabling logical reasoning above. The joint use of this promising technology with text mining techniques opens the doors to the automatic modeling of knowledge extracted directly from unstructured text (combining Statistical AI with Symbolic AI). Most of the methodologies used in state-of-the-art ontology learning are based on solutions strongly dependent on the domain, often requiring large labeled datasets and using machine learning models with a lack of explainability. This chapter presents an original unsupervised methodology for knowledge graph learning, starting from the descriptive text mining method already discussed.

5.1 Introduction

As mentioned in 3.7, in 2019 Gartner places *knowledge graphs* (KGs) among the innovation triggers on which higher expectations lie, forecasting a peak in 5-10 years. This growth in popularity is also accompanied by the evolution of graph neural networks, which in a few months have become a subject of great attention in the scientific world [188]. The high availability of unstructured data on the web has made the automatic acquisition of ontologies and knowledge graphs from text an important research area. In fact, handcrafting big knowledge representations is an extremely intensive and time consuming process, and it is impossible for all domains. *Ontology learning* (OL) studies the mechanisms and processes to transform heavy tasks like creation and maintenance of ontologies, into a semi or complete automatic process. Knowledge Extraction (KE) has become key to the Semantic Web, but interest in OL is not new (see e.g. [189], which dates back to 2001). It can be seen as a reverse engineering task, where domain model is reconstructed from input text by exploiting the formal structure saved in author's mind [190] (Figure 5.1).

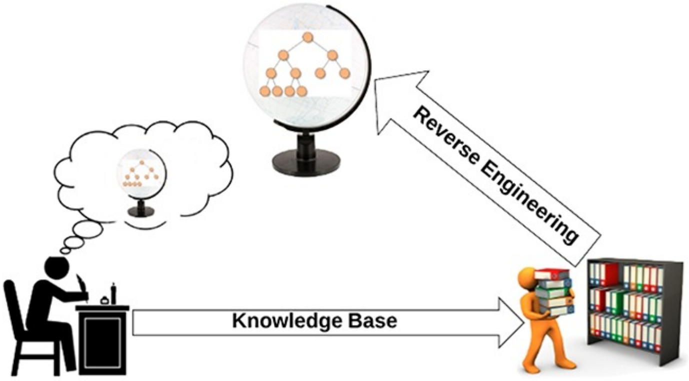
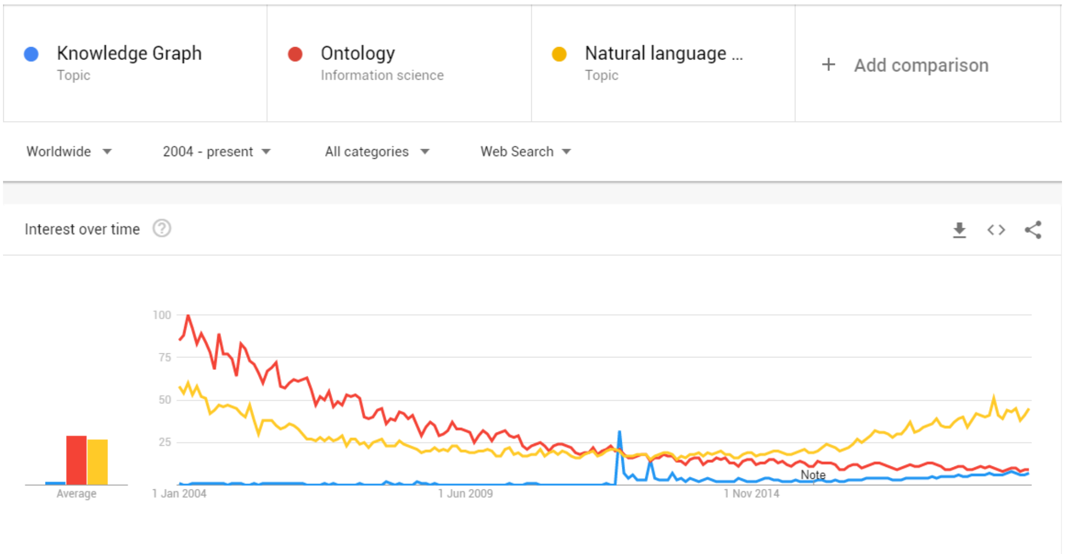


Figure 5.1: Ontology learning from text: a reverse engineering task
Obtained from [190]

One of the greatest challenges is therefore represented by the extraction of entities, relationships and knowledge in general from natural language. In many respects, the evolution of ontology learning goes hand in hand with that of *text mining*, and more specifically with that of *natural language processing* (NLP) [190]. This is also observable from the trends provided by Google for these concepts (Figure 5.2). Therefore, ontology learning is a multidisciplinary task that involves techniques of various fields. The union of *Statistical AI* and *Symbolic AI* (aimed at combining the advantages of both) is becoming increasingly frequent in terms of research, and represents a winning approach for the resolution of complex problems [13].
Knowledge modeling mainly refer to *descriptive* text mining (and not *predictive* one). In fact, in this case the objective is typically to represent knowledge that explains a certain phenomenon (and not that estimate the likelihood of a future outcome).

Several attempts have been made to bring some level of automation in the process of ontology acquisition from unstructured text, adopting several techniques to deal with various sub-problems. However, the proposed methodologies often require human intervention. Moreover, they are usually based on solutions



● Knowledge Graph
Topic

● Ontology
Information science

● Natural language ...
Topic

+ Add comparison

Worldwide ▾ 2004 - present ▾ All categories ▾ Web Search ▾

Interest over time

Figure 5.2: Interest over time for knowledge graphs, ontologies and natural language processing. The spread of ontologies has dropped significantly over years. In contrast, after the announcement from Google in May 2012, the attention toward knowledge graphs is constantly growing (with a positive correlation to the progress of the natural language processing).
Obtained with https://trends.google.com/trends/

strongly dependent on domain, or use tools that are ineffective in non-general contexts. As already noted in Chapter 4, large labeled datasets are often required to obtain better results. The integration of modern approaches (such as deep neural networks) can lead to advantages but their black-box nature does not make them particularly suitable for ontological modeling (where correctness, comprehension and reliability are required). Work on Explainable Artificial Intelligence (XAI) is still in a very early stage. Considering that descriptive text mining is a topic poorly covered in literature, new explanation-based methods and improvements in NLP techniques need to be merged into ontology learning systems for better results and performance.

Starting from the contribute described in Chapter 4, we propose an original methodology for unsupervised knowledge graph learning.

This methodology is highly flexible and enriches those already discussed in literature. It requires a truly negligible quantity of information compared to most other approaches, making it particularly suitable in contexts where there is no labeled data available.

5.2 Related Work

5.2.1 Ontology Learning Techniques

OL techniques have been widely investigated for various domain purposes and application scenarios. The whole process of ontology acquisition can be divided into several distinct steps, forming the so-called "ontology learning layer cake" [191] (Figure 5.3). This divide and rule vision arises from the desire to reuse and take advantage of the work already done in text mining and NLP fields. The process includes the extraction of terms and their synonyms from the underlying text, the combination of them to form concepts, the identification of taxonomic and non-taxonomic relationships between the found concepts, and finally the generation of rules. Each stage is dependent on results of the previous stage. So, one of the motors that drive OL itself is the recognition of patterns in the data that could originate new knowledge to further evaluation [53]. The main problem consists in understanding the meaning of concepts and related semantic relationships in an automated form, starting from the limited amount of information contained in a single document (especially in a social context such as that examined in this thesis).

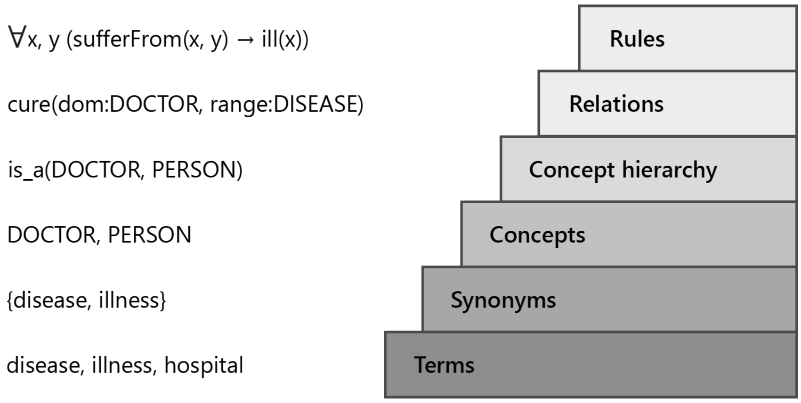


Figure 5.3: Ontology Learning Layer Cake

The papers published in literature can be distinguished on the basis of how they deal with the various stages. A great categorization has been proposed in a recent survey [190], which recognizes the presence of three classes of techniques: linguistic, statistical and logical. These are usable at different levels of the ontology learning layer cake. Although several researches have provided approaches for semi-automatic methodologies, the automatic learning of ontologies still requires a lot of work.

An efficient preprocessing of data using good linguistic techniques (such as part-of-speech tagging, dependency parsing and lemmatization) is often necessary to have higher accuracy [192].

Terms and concepts are generally captured by extracting sentence-level syntactic structures (such as noun phrases or verb phrases). More complex terms can be detected by searching for those that contain hypernyms [193]. Others solutions are based on a subcategorization frame approach, which is a linguistic concept that aims to recognize the words selected by another word when it appears in a sentence [194, 195]. Seed words research is another common linguistic technique focused on the extraction of domain-specific terms [196, 197, 198]. Regular expression patterns (sometimes with weightings) are also very popular [199]. Linguistic filters (based on part-of-speech patterns) can be used to get composite terms [199]. From a statistical point of view, C/NC value is used for multi-word terminology extraction [200, 201], and contrastive analysis can be applied for the removal of terms not relevant to the domain (also adopting domain relevance and domain consensus measures) [202]. Similarly, co-occurrence analysis and latent semantic analysis can be very useful for the extraction of concepts and the identification of implicit associations between them [203, 204, 205]. Clustering can also be used as an unsupervised learning approach for this sub-task [206]. Moreover, named entity recognition systems are a great way to obtain interesting terms, and can be implemented with linguistic grammar-based techniques or with statistical models [207].

The techniques proposed in literature for the construction of a hierarchy of concepts use term subsumption (algorithms based on equations with conditional probabilities) [208], formal concept analysis (where the interconnections are focused on the attributes shared by the objects) [209] and hierarchical clustering [210, 211, 212].

A simple class hierarchy with only “is_a” relations is not sufficient. To understand the meaning of a class or what is a class, the definition of its properties is also required. In the end, it is the way in which a concept relates to other knowledge to define the meaning or semantics. Relation extraction as a part of ontology generation and population is a challenging task, and has been pursued for more than a decade. An approach widely used in the ontological field for the discovery of non-taxonomic relationships is association rule mining, an unsupervised data mining method based on finding rules to predict the co-occurrences of elements [209, 213, 214]. There are also supervised machine learning algorithms, requiring training examples for relation extraction from domain text: bootstrapping methods (weekly supervised) [215], logistic regression methods [216] etc. A simpler alternative is to use regular expressions, patterns and rules crafted by human experts to identify relationships of various types (e.g., synonymy, hyponymy, causality) between concepts [207, 199]. A pattern learning for relation extraction using FreeBase is presented in [217]. Open information extraction (OIE) systems identify the relations between two terms by extracting and analyzing the text(s) occurring between these terms in the corpus. They are based first on the search for potential groups of words related to each other (using statistical approaches based on frequency distribution, or semantic approaches focused on the WordNet path similarity measure), and then on the extraction of text patterns [199].

The solution mainly adopted at the final stage of ontology learning layer cake is inductive logic programming, a subfield of machine learning that uses first-order logic to derive hypotheses based on background knowledge [218].

A summary of the techniques typically used at the various levels of the ontology learning layer cake is shown in Table 5.1. In [190], after reviewing 140 papers, the authors observed that a hybrid approach comprising of both linguistic and statistical techniques produces better ontologies.

Step #	Step Description	Techniques		
		Linguistics	Statistical	Logical
0	Pre-processing	Part of Speech Tagging Dependency Parsing Lemmatization		
1,2,3	Term/Synonym/ Concept Extraction	Regular expressions Syntactic Analysis Linguistic filters Subcategorization Frames Seed words extraction Named Entity Recognition	Named Entity Recognition Term Weighting C/NC value Contrastive Analysis Co-occurrence analysis Latent Semantic Analysis Clustering	
4	Concept Hierarchy	Regular expressions Dependency Analysis Lexico Syntactic Pattern WordNet-based	Term Subsumption Formal Concept Analysis Hierarchical Clustering	
5	Relation Extraction	Regular expressions Open Information Extraction WordNet-based	Association Rule Mining Bootstrapping Logistic Regression Open Information Extraction	
6	Rules and axioms			Inductive Logic Programming

Table 5.1: Summary of Ontology Learning Techniques

It is important to be careful not to use rigid techniques that require changes in the algorithm or data to support knowledge extraction from future concepts, now unavailable and unknown (e.g., new medical treatments). Dictionary-based methods lack in accuracy after several years because of the arrival of new entities to the field. Especially in the medical domain, there is a higher necessity for knowledge bases capable of dynamically evolving over time and with data. New biomedical entities (such as diseases and symptoms) are frequently introducing. In [207] it is presented an approach for the automatic update of the numerical probability values related to each relationship, evolving the ontology with the arrival of new patient entries.

In conclusion, many of the available techniques are supervised (require labeled datasets for training) and highly domain-specific (especially with regards to patterns and dictionaries). Therefore, their applicability in other domains is generally difficult, if not impossible. Some disregard the frequent grammatical errors in text. The integration of the coreference resolution for the management of pronouns can in several cases improve the results in terms of recall. A manual inspection is often necessary to remove irrelevant concepts or relationships between the extracted ones. In fact, determining suitable thresholds to automate this process is difficult [199]. WordNet-based solutions have the limitation of being able to manage only elements included in the supported synsets (sets of cognitive synonyms consisting of nouns, verbs, adjectives and adverbs). Furthermore, composite terms cannot be processed using WordNet, as they are not managed by synsets [199].

5.2.2 Ontology Learning Tools

Several tools for OL were born over time and the best known are listed below.

- *TERMINAE* is a tool to facilitate the creation of domain ontologies. Uses linguistic and knowledge engineering to extract terms, synonyms, taxonomic and non-taxonomic relations. With regard to the state-of-the-art, it is one of the most supervised methods in the trend of ontology learning [219]. Although it has been updated several times, it was first designed about 20 years ago.

- *Text2Onto*[1] is the official successor of *TextToOnto*, a framework for supervised ontology learning from text. Developed at the University of Karlsruhe, it is considered an hybrid tool as it exploits both linguistic and statistical techniques in order to extract terms and relations from underlying corpus. Text2Onto uses GATE 4.0[2] for sentence detection, tokenization, POS-tagging and application of pattern rules. It is based on Probabilistic Ontology Model (POM), i.e. all learned objects are enhanced by calculated probabilities in such manner that a user can decide whether to include this object into the ontology or not. It implements some text mining algorithms on textual data and includes several data

[1] `http://neon-toolkit.org/wiki/1.x/Text2Onto.html`

[2] An open source software toolkit capable of solving several text processing problems. `https://gate.ac.uk/`

processing: concept extraction (with relative term frequency, TF-IDF, Entropy, C/NC), concept inheritance (Wordnet and its hypernym network), relation extraction (lexico-syntactic patterns), equivalence (context similarity) [220].

- *OntoGen*[3] is an editor focused on the use of text mining and machine learning techniques for the extraction of terms, concepts and taxonomies [221]. Supports the visualization of concepts.

- *CRCTOL* is a semantic-based domain ontology learning system. It uses statistical lexico-syntactic association rules to extracts concepts, taxonomies and non-taxonomic relations. At the component level, quantitative evaluation by comparing with Text-To-Onto and its successor Text2Onto has shown that CRCTOL is able to extract concepts and semantic relations with a significantly higher level of accuracy [222].

- *DBpedia Spotlight*[4] is a tool for automatically annotating mentions of DBpedia resources in text. It is available as a REST service, or downloadable [223].

- *OntoPop* is an ontology population system based on LSA[224].

- *FRED*[5] is a tool for automatically producing RDF/OWL ontologies and linked data from text. The method is based on deep semantic parsing. It leverages Natural Language Processing components for performing Named Entity Resolution Coreference Resolution, and Word Sense Disambiguation [225].

The main objective of almost all of these tools is not to automatically build an ontology by learning on a body of texts, but help user to do it. According to [207], dynamic ontology construction is a complicated and tedious process using existing solutions such as Text-to-Onto and Text2Onto. These systems detect relations using association rule mining and predefined regular expressions to expand the ontology dynamically. As a result, they are not very flexible and it is difficult to use them for extracting domain specific concepts. Text2Onto is one of the most mentioned OL tools in academic documents [190], but together with OntoGen it is frequently criticized for his great limitations [226]. FRED stands out for its functionality, but does not provide the possibility to manage and customize the semantics of relationships (limiting its applicability). A comparison of other KE tools can be found in [227]. Finally, a very interesting

[3] `http://ontogen.ijs.si/`

[4] `http://dbpedia-spotlight.github.com/demo`

[5] `http://wit.istc.cnr.it/stlab-tools/fred`

product is Ontotext Platform[6], which provide an enrichment suite for text mining and semantic annotation.

5.3 A new Knowledge Graph Learning Methodology

In this section we propose an original methodology for the formalization of ontologies or knowledge graphs starting from the results obtained by means of descriptive text mining solutions. More specifically, we show how it is possible to automatically represent the knowledge extracted with the mixture of linguistic and statistical techniques adopted for the unsupervised method presented in Chapter 4. So, the proposed knowledge graph learning methodology inherits all the advantages of the previous contribution, especially in terms of flexibility and modularity.

5.3.1 Preprocessing

This step is directly included in the descriptive text mining method. It generally concerns NER handling, application of filters, quality preprocessing (noise removal), lemmatization, documents classification, analysis preprocessing, term-document matrix construction, and term weighting. However, the specific transformations vary according to the domain, the characteristics of the data available, and the objective of the analysis.

5.3.2 Term/Synonym/Concept extraction

After the labeling of the documents with the tags linked to the results of the Named Entity Recognizer, the terms (word-level unigrams with a frequency greater than a minimum threshold) are distinguished between standard and entity ones. In the field of ontology/knowledge graph learning, the definition of *term* can be compared to that of standard term in the case of descriptive analysis. Similarly, a *concept* corresponds to an entity recognized by the NER system (for which types, class and optional links to external knowledge bases are known). Terms linked by high cosine similarities in the latent semantic space generated through the LM can be considered *synonyms*. Consequently, the latter can be identified through clustering operations.

[6] `https://www.ontotext.com/`

5.3.3 Concept hierarchy

The adoption of a NER based on hierarchical types or the prior definition of a hierarchy capable of extending in this sense the implementation of a certain NER system, also respond to the need to create a taxonomy of concepts. It is therefore possible to extract is_a relationships from the flat results provided by a Named Entity Recognizer.

5.3.4 Relations

In addition to hierarchical relationships, it is possible to model the correlations between the terms emerged with a global analysis. By making use of rules or by training a classifier, it is also possible to extract/infer more specific relations from correlation ones. For example, in the case of a strong correlation between a medical treatment and a specialized center, the relation type "has_correlated_term" can be replaced with "is_performed_at". Moreover, by launching local analyzes on specific concepts (requiring only the corresponding regex), it is also possible to represent the membership relationships of a term to its class descriptions.

The relationships extracted with text mining techniques are not certain, but probabilistic. The use of cosine similarity remapped on the interval $[0, 1]$ or of the p-value associated with the chi-squared test relative to the term-term/term-class correlation, allows access to this probability value (directly available in the case of P-LSA). This methodology therefore proposes the labeling of relations with the respective probability (as already discussed in 3.6).

5.3.5 Linked Open Data

The use of NER systems returning the references to external knowledge bases for the tagged entities or the application of tools for the pursuit of this purpose, enables the connection with other controlled vocabularies (e.g., DBpedia, GeoNames).

In conclusion, this hybrid methodology is able to identify and properly define a set of relevant concepts that characterize a given application domain, capturing also the semantic relationships among them from textual data. Depending on the design choice, this methodology can lead to an ontology (if the terms are modeled as classes and relations as an object property) or to a knowledge graph (if the terms are modeled as individuals and the relationships managed with OWL2 punning[7]).

[7] Technique focused on creating an individual with the same IRI as a class.

Chapter 6

Case Study

This chapter shows the application of the contributions to a case study in the medical field, specifically focused on the domain of rare diseases. Experiments are conducted on unlabeled conversational posts about a rare auto-immune disorder, called “Esophageal Achalasia”. With the collaboration of *Associazione Malati Acalasia Esofagea (AMAE) Onlus*[1], the main Italian association for this disease, a descriptive analysis is applied with the aim of collecting useful information to improve patients’ living conditions. In particular, we demonstrate how it is possible to obtain scientific medical correlations directly from the large set of short dialogue messages shared on social networks. Through the discovery of statistically significant evidences from the data, we also learn a knowledge graph in an unsupervised way (by extending the disease node on ORDO with a patient-centered vision).

6.1 Esophageal Achalasia Overview

This section provides a brief overview of the disease, so that the results of the analysis can then be better appreciated.

6.1.1 Disease description

Achalasia (*ORPHA:930*) is a rare disorder of the esophagus. It has an annual incidence[2] of approximately 1/200,000 to 1/59,000 and a prevalence[3]

[1] `http://www.amae.it/`

[2] A measure of the probability of occurrence of a given medical condition in a population within a specified period of time. It conveys information about the risk of contracting the disease, and it refers to new cases.

[3] The proportion of cases in the population at a given time. It indicates how widespread the disease is.

rate estimated to be 1/10,000 [228].

Esophageal Achalasia is characterized by impaired ability to push food down toward the stomach (peristalsis). This problem is due to the failure of the ring-shaped muscle at the bottom of the esophagus, the lower esophageal sphincter (LES), to relax (Figure 6.1). In fact, the movement of food through the tube is made possible precisely by the contraction and relaxation of the sphincter [229]. The slow decompensation of the muscular layer causes severe dilatation in the long term, which is a characteristic of megaesophagus.

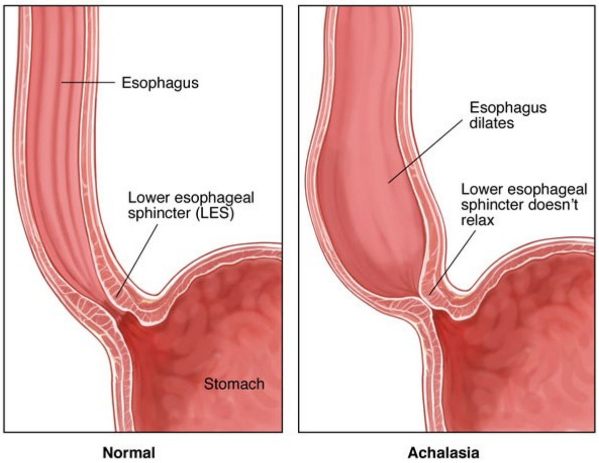

Esophagus
Lower esophageal sphincter (LES)
Stomach
Normal
Esophagus dilates
Lower esophageal sphincter doesn't relax
Achalasia

Figure 6.1: Constant contraction of the LES in patients with Esophageal Achalasia

Obtained from `https://anatomy-medicine.com/gastroenterology/219-achalasia.html`

Achalasia is a heterogeneous disease categorized into 3 distinct types based on manometric patterns: type I (classic) with minimal contractility in the esophageal body, type II with intermittent periods of panesophageal pressurization, and type III (spastic) with premature or spastic distal esophageal contractions [230] (Figure 6.2).

Although the precise etiology is unknown, it is often thought to be either autoimmune, viral immune, or neurodegenerative. Some familial cases have been reported, but the rarity of familial occurrence does not support the hypothesis that genetic inheritance is a significant etiologic factor. Esophageal Achalasia has been associated with viral infections and auto-antibodies against myenteric plexus have been found, but the casual relationship remains unclear [228].

The main symptoms of achalasia are dysphagia (difficulty in swallowing), regurgitation of undigested food, chest pain behind the sternum, and weight loss. Disphagia tends to become progressively worse over time and involve both

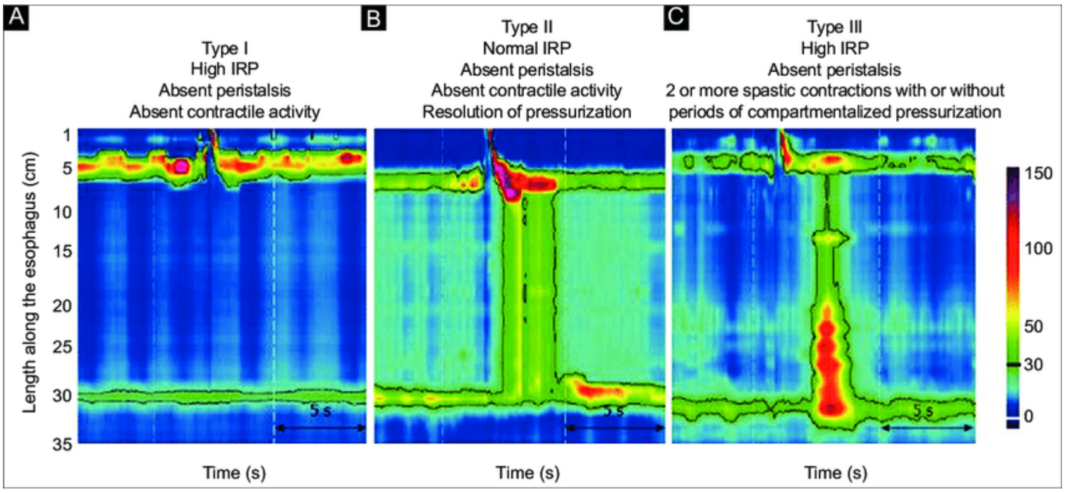

A
Type I
High IRP
Absent peristalsis
Absent contractile activity
B
Type II
Normal IRP
Absent peristalsis
Absent contractile activity
Resolution of pressurization
C
Type III
High IRP
Absent peristalsis
2 or more spastic contractions with or without
periods of compartmentalized pressurization
Length along the esophagus (cm)
1
5
10
15
20
25
30
35
5 s
Time (s)
150
100
50
30
0

Figure 6.2: High-resolution manometry showing the 3 subtypes of Esophageal Achalasia

Obtained from [230]

fluids and solids. Some people may also experience coughing when lying in a horizontal position. The chest pain experienced, also known as cardiospasm and non-cardiac chest pain can often be mistaken for a heart attack. It can be extremely painful in some sufferers. Food and liquid, including saliva, are retained in the esophagus and may be inhaled into the lungs (aspiration) [231].

Due to the similarity of symptoms, achalasia can be mistaken for more common disorders such as gastroesphageal reflux disease (GERD), hiatus hernia, and even psychosomatic disorders.

Specific tests for achalasia are barium swallow and esophageal manometry. In addition, endoscopy of the esophagus, stomach, and duodenum (esophagogastroduodenoscopy or EGD), with or without endoscopic ultrasound, is typically performed to rule out the possibility of cancer [231].

Treatment is to ease the symptoms of achalasia by decreasing the outflow resistance caused by a non-relaxing and hypertensive lower esophageal sphincter. Current treatment modalities for achalasia are nonsurgical (pharmacotherapy, endoscopic botulinum toxic injection, and pneumatic dilatation) or surgical (laparoscopic heller myotomy, abbreviated LHM, and peroral endoscopic myotomy, abbreviated POEM). Pharmacologic treatments provide only short-term relief of symptoms, and are primarily reserved for patients who are waiting for or who refused more definitive therapy or surgery. Endoscopic injection of botulinum toxin can be used in high-risk patients or those who relapse after myotomy. Pneumatic dilatation of the esophagus via endoscopy is the most cost-effective nonsurgical therapy for achalasia. Is the first treatment option for a patient in whom surgery fails. The recommended step for reducing pressure across the lower esophageal sphincter is surgical myotomy. This procedure will cut the circular muscle fibers running across the lower esophageal sphincter, leading to relaxation. LHM can potentially cause uncontrolled gastroesophageal reflux, so it typically pairs with an anti-reflux procedure such as Nissen, the posterior (Toupet), or the anterior (Dor) partial fundoplication. The anterior fundoplication is the most common choice. POEM is an effective minimally invasive alternative to laparoscopic Heller myotomy to treat achalasia at limited centers. Dissection of the circular fibers of the LES is achieved endoscopically, leading to relaxation of the LES; however the risk of gastroesophageal reflux is high because it does not include an antireflux procedure. Esophagectomy is the last resort [232].

6.1.2 Patient Organisations

In Italy there are only two patient organisations for Esophageal Achalasia [233]:

- AMAE - Associazione Malati Acalasia Esofagea ONLUS;
- ALMA - Associazione Libera Malati Acalasia e altre malattie dell'esofago ONLUS.

AMAE

Born over fifteen years ago at the behest of a group of patients, AMAE has as principal statutory purposes the representation of the instances of subjects affected by the rare pathology, the support to family caregivers, and the construction of proactive interactions with the health professionals involved in the care path. The actions carried out by the association are therefore aimed at conveying a correct and usable knowledge to the community of patients and carers, hoping to obtain a significant social impact and a real improvement in the rare sufferer's quality of life.

The project presented in this thesis fits perfectly with AMAE's research objectives. For this reason, a collaboration has immediately been reached, in the interests of researchers, patients and their caregivers.

6.1.3 Patient Communities

The main patient communities on the Web for Esophageal Achalasia can be found on Facebook. Although in the past ad-hoc platforms have been created for the dialogue between rare patients, many of these have fallen into disuse, or are sparsely frequented and poor in content compared to the most popular social networks. The community dedicated to Achalasia on RareConnect[4], for example, at the time of writing has only 5 members and 7 posts.

Facebook Groups, in particular, are the most popular communication tool. Although not specifically designed for this, they benefit from one of the largest user bases in the world, have a high margin of personalization, and are maintained with high frequency.

There are two Italian groups dedicated to Esophageal Achalasia.

- *Acalasia esofagea... I malati "rari" non sono soli...!*[5]
  Creation Date: October 16, 2008
  Number of users: ≈ 2000
- *Acalasia esofagea: pazienti a confronto*[6]
  Creation Date: March 4, 2015
  Number of users: ≈ 900

They are managed respectively by AMAE and ALMA.

AMAE Facebook Group

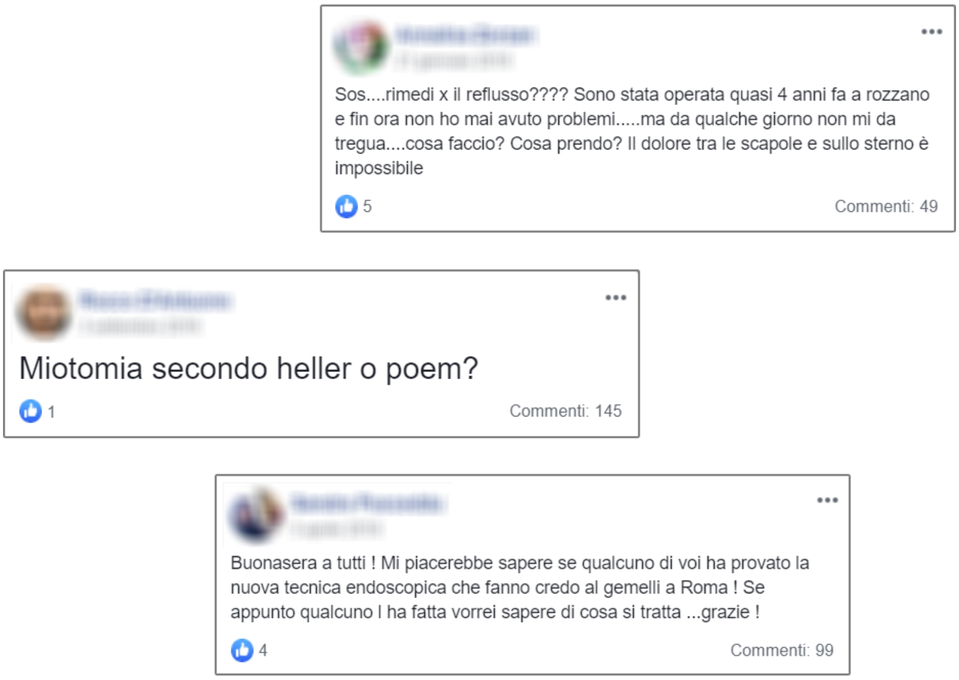
Acalasia esofagea... I malati "rari" non sono soli...! is the group with the largest amount of data. Being the one directly handled by AMAE, it also constitutes the source for the creation of the dataset. Thanks to its private nature, administrators exercise control on the quality of published posts (which need approval). This filtering operation is very valuable as it strongly limits the presence of fake news and in general of harmful content towards other patients. As a result, this also prevents these data from compromising the results of the analysis. A large part of the users are patients and doctors. Many issues are discussed within the group. People discuss the symptoms felt, their experiences, the effectiveness of the treatment carried out, the best intervention technique, the foods to avoid and those recommended instead, the most accredited medical centers and doctors, drugs, exemptions, etc. Some examples of posts extracted directly from the group are shown in Figure 6.3. The idea is to take advantage of the large amount of data accumulated over time, offering an original research contribution. Creating a system for the automatic extraction of knowledge contained within posts and comments, and for the relative representation of it, can have a strong social impact. Patient opinions are very valuable for researchers and pharmaceutical companies, but also for doctors and patient organizations. Furthermore, having direct access to what has been expressed on a certain topic by all the other users is of incredible utility even for the

[4] https://www.rareconnect.org/en/community/achalasia

[5] https://www.facebook.com/groups/36705181245/

[6] https://www.facebook.com/groups/413624182145929/

patient himself (who is able to obtain a targeted response in a short time, statistically proven on a high number of documents). In the literature there are no automated solutions to meet these needs.



Sos....rimedi x il reflusso???? Sono stata operata quasi 4 anni fa a rozzano e fin ora non ho mai avuto problemi.....ma da qualche giorno non mi da tregua....cosa faccio? Cosa prendo? Il dolore tra le scapole e sullo sterno è impossibile

5 Commenti: 49

Miotomia secondo heller o poem?

1 Commenti: 145

Buonasera a tutti ! Mi piacerebbe sapere se qualcuno di voi ha provato la nuova tecnica endoscopica che fanno credo al gemelli a Roma ! Se appunto qualcuno l ha fatta vorrei sapere di cosa si tratta ...grazie !

4 Commenti: 99

Figure 6.3: Examples of posts shared by patients in *Acalasia esofagea... I malati "rari" non sono soli...!*

6.2 Patient Survey

Understanding what patients want and need from rare disease research and data sharing is important to ensure their participation and engagement in the process, checking that these wishes and needs are embedded within research itself [8]. In order to be able to direct the analysis towards obtaining information as useful as possible for the community, we first chose to administer a survey to the users registered in the group. More in detail, 282 responses were collected on the degree of interest (released as a score in the range $[0, 10]$) in various research topics, distinguishing between patients (252) and caregivers (30). The results are shown in Figure 6.4. As can be seen, many areas of research have a high interest and these values are often similar to each other (to justify the sufferers' need for answers). The subjects on which patients require more investigations are, in order, *medical treatments*, *diagnoses*, *symptoms* and *medical centers*. The opinions expressed by the caregivers do not differ much.

In both cases, gestation and drugs are the least voted topics. Having this data is important for understanding what are the requirements that the analysis must try to satisfy and what are the best technological tools to do it.

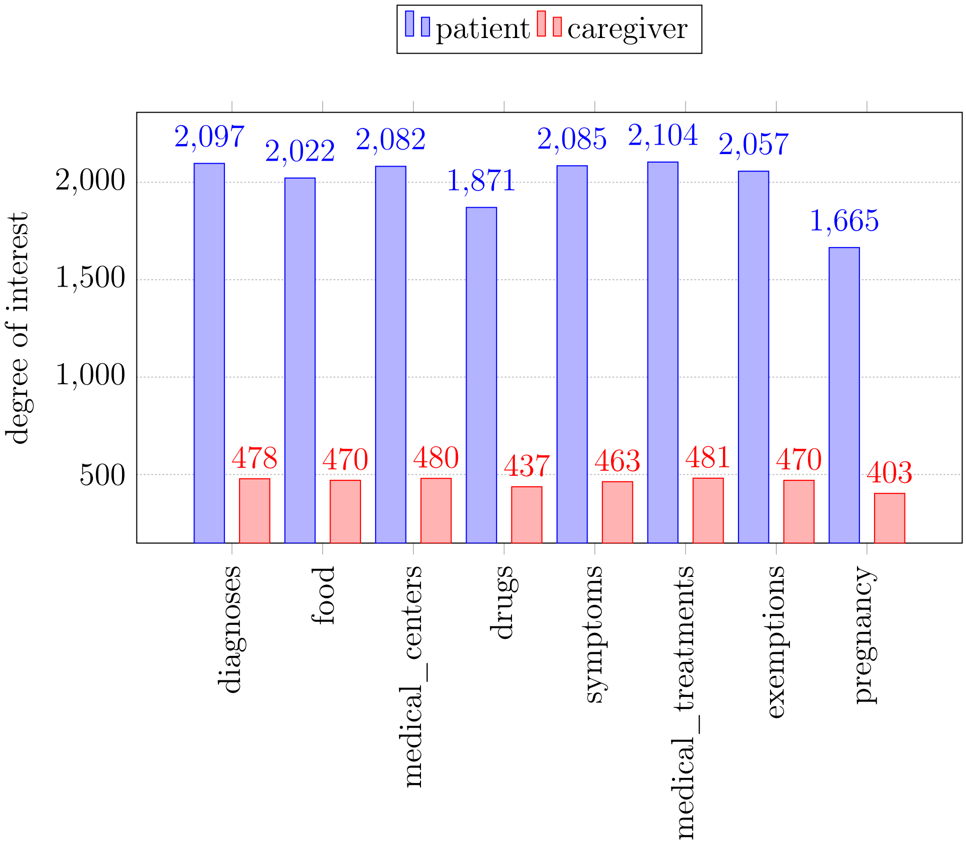


Figure 6.4: Interest in the main research areas related to Esophageal Achalasia, expressed by patients and caregivers

6.3 Preliminar Technical Choices

R and **Python** are the two most popular programming languages for data science. From a systemic point of view, we chose to use them together, in order to combine the strengths of both. In fact, they have advantages and disadvantages depending on the specific task to be faced. For example, R has a strong memory optimization and is able to manage extraordinarily well sparse matrices with less than 1% of cells containing information (a frequent requirement in genomics contexts). Python is popular for its speed and most code-side functionality due to the general-purpose nature of the language. R is widely considered the best tool for making beautiful graphs and visualizations. Moreover, Python is great for mathematical computation, but it doesn't have as many libraries as R, and there are no module replacement for the hundreds of essential R packages. Finally, learning both of them is the ideal solution.

To combine and make the two worlds work together, we used **Reticulate**[7] (a package which provides a comprehensive set of tools for interoperability between Python and R). It includes facilities for translating objects, and calling Python from R in a variety of ways (e.g., R Markdown). A Python session is therefore embedded within a R session, offering the possibility for developers to never leave RStudio as a development environment.

6.4 Dataset Construction

6.4.1 Posts and Comments

Most modern social systems allow some form of access to their data through an Application Programming Interface (API), often accompanied by additional utilities such as keyword and date filters. Automatic retrieval of posts and comments can be achieved with the **Facebook Graph API**[8], the primary way to get data into and out of the Facebook platform. Although the Graph API is HTTP-based, access to it by means of R is simplified thanks to the **RFacebook**[9] package. A preliminary step for the use of the functions contained in this package is the creation of a new Facebook App (Figure 6.5). Once the application is provisioned and configured appropriately, it is possible to connect to it from a R session and perform authentication. A test user with the necessary permissions has been created for the project, the most important of which is *groups_ access_ member_ info*.

The two main functions provided by RFacebook that have been used are:

- `getGroup`, for the acquisition of the Facebook group posts;
- `getPost`, for finding all the information associated with a specific post (including the list of comments).

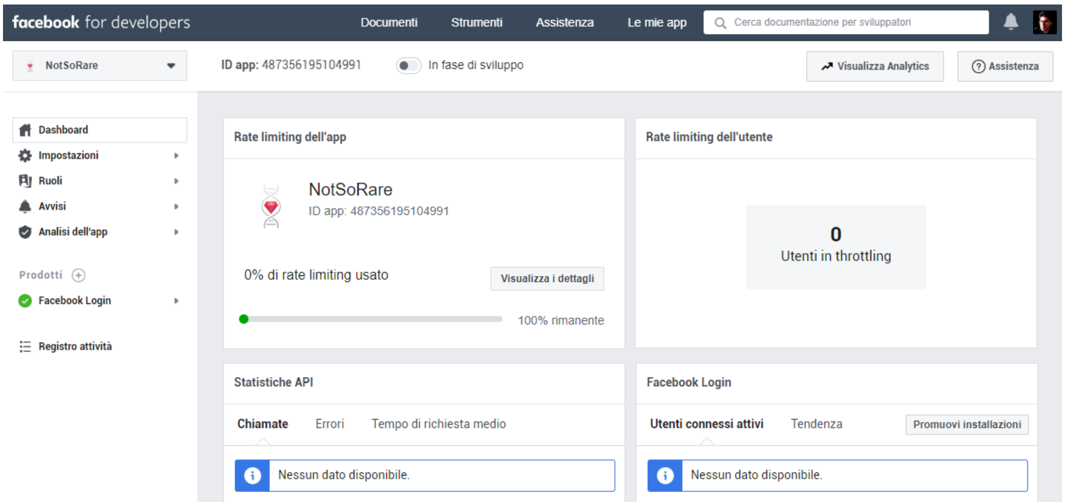
It is important to note that RFacebook is no longer maintained and issues are not monitored, as indicated in the official repository[10]. The security updates

[7] https://rstudio.github.io/reticulate/

[8] https://developers.facebook.com/docs/graph-api/

[9] http://cran.r-project.org/web/packages/Rfacebook

[10] https://github.com/pablobarbera/Rfacebook



ID app: 487356195104991

In fase di sviluppo

NotSoRare
ID app: 487356195104991

0% di rate limiting usato

100% rimanente

0
Utenti in throttling

Nessun dato disponibile.

Nessun dato disponibile.

Figure 6.5: Dashboard of a Facebook App

of the Facebook API in April 2018[11] made many libraries obsolete and incorrect. For this purpose, the code of the files hosting the two previously mentioned functions has been completely rewritten.

By carrying out the download in compliance with the threshold limits, two CSV files were created containing respectively all the data of the posts and first-level comments. Specifically, the data refer to 6,917 posts and 61,692 comments, published between 21/02/2009 and 05/08/2019.

6.4.2 Named Entity Recognizer

Given the high amount of data and the intention to carry out multiple analyzes, NER is not applied on every occasion. It is therefore performed on the original documents and its results stored on files, expanding the dataset and providing the possibility of being able to use them when necessary.

Considering the heterogeneity and specificity of the entity types linked to the topics of interest from the patient's point of view, it is necessary to adopt a NER system sufficiently flexible and expressive to allow the pursuit of the objectives.

TextRazor[12] is an excellent solution for this task. It is a technology making use of state-of-the-art NLP and AI techniques to parse, analyze and extract semantic metadata from textual content. In particular, TextRazor offers a complete cloud or self-hosted text analysis infrastructure, taking advantage of a huge knowledgebase of real-life facts to help rapidly extract the value from

[11] `https://developers.facebook.com/docs/graph-api/changelog/`

[12] `https://www.textrazor.com/`

documents. It supports 12 languages, including Italian (ensuring compatibility with the case study). Although TextRazor's API is able to manage different components, for the purposes of this application case we focus only on that of Entity Recognition. TextRazor achieves industry leading NER performance by leveraging the knowledgebase already mentioned (extracted from various web sources, including Wikipedia, DBPedia and Wikidata). The system uses a matching engine to perform a quick lookup in the text, starting from a dictionary made up of millions of different possible entities. At present (model version: 2019-11), the known entities are 28,037,058. Consequently, it is the most comprehensive Named Entity Recognizer to the best of our knowledge. Furthermore, it takes into account disambiguation, thanks to a deep understanding of the context. It also combines a number of different signals into a single confidence score for each entity, which ranges from 0.5 to 10 (with 10 representing the highest one). Finally, TextRazor has two fundamental characteristics for the proposed descriptive text mining method:

- it can identify thousands of taxonomy types[13] in both Freebase and DBPedia;
- it disambiguates and links entities to canonical IDs in the linked web (returning where possible a link to Wikipedia, the disambiguated Freebase ID, and the disambiguated Wikidata QID).

To take full advantage of the data returned by TextRazor's Named Entity Recognizer, therefore, it is necessary to know the difference between *DBpedia* and *Wikidata*. They fulfill very different tasks. DBpedia extracts structured data from the infoboxes in Wikipedia, and publishes them in RDF and a few other formats. Wikidata on the other hand provide a secondary database of structured data that everyone can edit. So, instead of extracting structured data from infoboxes, it will allow infoboxes to be created from structured data [234].

Figure 6.6 shows an example of Entity Recognition component, with the illustrated functionalities.

In conclusion, TextRazor's REST API is used to complete the NER task on the textual component of the posts and comments previously downloaded. The results of this operation are also stored on CSV files, and concern 15,687 entities for posts and 73,095 for comments.

Thanks to Toby Crayston, founder of TextRazor, for facilitating the development of this academic project by increasing free limits, and for showing enthusiasm in its realization.

[13] `https://www.textrazor.com/types`

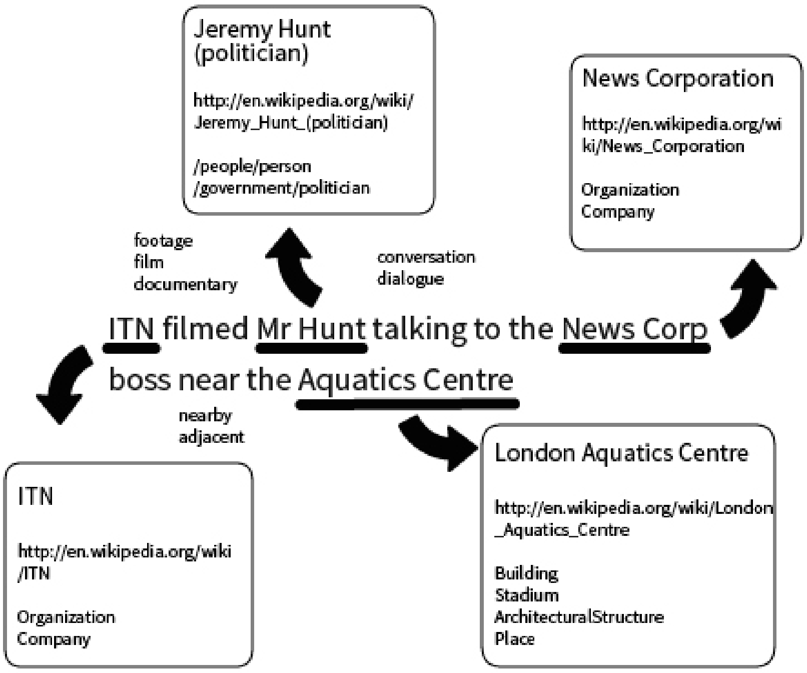

Jeremy Hunt (politician)
http://en.wikipedia.org/wiki/Jeremy_Hunt_(politician)
/people/person
/government/politician
News Corporation
http://en.wikipedia.org/wiki/News_Corporation
Organization
Company
footage
film
documentary
conversation
dialogue
ITN filmed Mr Hunt talking to the News Corp boss near the Aquatics Centre
nearby
adjacent
ITN
http://en.wikipedia.org/wiki/ITN
Organization
Company
London Aquatics Centre
http://en.wikipedia.org/wiki/London_Aquatics_Centre
Building
Stadium
ArchitecturalStructure
Place

Figure 6.6: TextRazor's Entity Recognition with disambiguation and linking to other knowledgebases

Obtained from https://www.textrazor.com/

6.4.3 Dataset Structure

The original dataset is made up of 4 CSV files, for a total size of 36.9MB. The structure of the dataset itself is shown below, describing the content and organization of each file inside it, as well as explaining the meaning of the various fields where necessary.

```
dataset
 ├─ posts_ita.csv
 ├─ comments_ita.csv
 └─ textrazor
     ├─ textrazor_ner_posts_ita.csv
     └─ textrazor_ner_comments_ita.csv
```

1. **posts_ita.csv**
   Size: 2,773 KB.
   Content: posts data.
   Fields: id *(the Facebook post identifier)*, message, created_time, updated_time, type *(a string indicating the object type of this post; es. status,*

photo, link), status_type *(description of the type of a status update; e.g. added_photo, shared_ story)*, is_published, is_popular *(indication of the popularity of the post, calculated by Facebook on the basis of the exceeding of a threshold by the set of actions carried out in terms of interaction)*, place, event, link, link_name, link_caption, comments_count, likes_count, love_count, wow_count, haha_count, sad_count, angry_count, thankful_count, shares_count.
Fields used for analysis: id, message, created_time.

2. **comments_ita.csv**
   Size: 14,053 KB.
   Content: first-level comments data.
   Fields: parent_post_id, id, from_id, from_name, parent, message, attachment, created_time, updated_time, likes_count, comments_count.
   Fields used for analysis: id, message, created_time.

3. **textrazor_ner_posts_ita.csv**
   Size: 3,997 KB.
   Content: entities extracted from posts throguh TextRazor's API.
   id *(progressive identifier locally to the single post for the recognized entity)*, dbpedia_types *(es. Person, Food, Place, Disease)*, matching_tokens *(list of token positions that granted the entity recognition)*, entity_id *(identity identifier from the localized version of Wikipedia with respect to the language of the document)*, freebase_types *(Freebase types list)*, confidence_score *(the confidence that TextRazor is correct that this is a valid entity, between 0.5 and 10)*, wiki_link *(link to Wikipedia for this entity)*, matched_text *(original textual component that granted the recognition of the entity)*, relevance_score *(relevance of the entity within the text, from a minimum of 0 to a maximum of 1)*, entity_english_id *(entity identifier in English Wikipedia)*, starting_pos, ending_pos, post_id, freebase_id *(entity identifier on Freebase)*, wikidata_id *(Wikidata QID for the entity)*, unit *(es. Number)*, crunchbase_id, lei, permid, figi, source_id, custom_entity_id *(identity of the entity, if belonging to a custom dictionary)*, data.entity_english_id, data.wiki_link, data.wikidata_id.
   Fields used for analysis: id, matched_text, entity_id, dbpedia_types, freebase_types, freebase_id, wikidata_id, wiki_link.

4. **textrazor_ner_comments_ita.csv**
   Size: 17,026 KB.
   Content: entities extracted from first-level comments through TextRazor's

API. It is structured in the same way as “textrazor_ner_posts_ita.csv” and the attributes used for the analysis are the same.

6.5 General aspects

This section briefly illustrates some general observations related to the application of the descriptive text mining method on the case study.

6.5.1 Text Razor NER Filtering and Mapping

As already noted, TextRazor’s Entity Recognition supports a large amount of data types in terms of labeling. However, three problems need to be addressed.

1. Entities could be labeled with a low confidence score. So, we would like to eliminate uncertain results.
2. Some terms may be labeled with types not of interest in the rare disease domain (e.g., /american_football/football_player, Asteroid, Bank, Motorcycle). Therefore, we would like to have the possibility to filter only entities having a significant type for analysis purposes.
3. The types assigned by the NER system follow two distinct taxonomies (Freebase and DBPedia). The Freebase hierarchy is much more detailed and structured than that of DBPedia, and therefore they are not directly comparable. Given an entity within a document, the Named Entity Recognizer is not necessarily able to associate a type of both hierarchies with it. In other words, a recognized entity could have either Freebase types, DBPedia types, or both. Maintaining this diversification is inconvenient within the text mining method. We would like to map the NER tags into a new and unified taxonomy.

Looking at the data, we saw how it is sufficient to remove the entities with minimal confidence to reduce noise. Consequently, only recognized entities with a confidence score greater than 0.5 are maintained. In addition, we only select entities that have at least one type of interest for analysis.

Then we define our own list of types, which we consider suitable for the domain and which we want to use during the analysis. Each type is associated with the corresponding set of Freebase and DBPedia tags. This allows a mapping from the results of TextRazor to our modeling. Specifically, each occurrence of an original type within an entity is replaced with the unified version, eliminating any duplicates. Table 6.1 shows the unification scheme used within the thesis.

Freebase Types	DBPedia Types	Unified Types
	Beer, Vodka, Wine	AlcoholicBeverage
/food/beverage, /food/beverage_type	Beverage	Beverage
/food, /biology/animal	Food, Animal	Food
	Species	Species
/chemistry/phase_of_matter		PhaseOfMatter
/location/it_comune /location/citytown	Settlement	City
/location/it_region	Region	Region
/location/country	Country	Country
/location/statistical_region	PopulatedPlace	Place
/medicine/hospital	Hospital	Hospital
/medicine/risk_factor		RiskFactor
/medicine/disease /medicine/infectious_disease	Disease	Disease
/medicine/disease_cause		DiseaseCause
/medicine/contraindication		Contraindication
/people/cause_of_death		DeathCause
/medicine/type_of_infectious_agent		InfectiousAgent
/medicine/symptom		Symptom
/medicine/icd_9_cm_classification		Icd9Classification
/medicine/condition_prevention_factors		PreventionFactor
/medicine/medical_treatment		MedicalTreatment
/medicine/anatomical_structure	AnatomicalStructure	AnatomicalStructure
/biology/protein	Biomolecule	Biomolecule
/medicine/drug_class		DrugClass
/medicine/drug	Drug	Drug
/medicine/drug_ingredient		DrugIngredient
/medicine/drug_formulation		DrugFormulation
/chemistry/chemical_compound /chemistry/chemical_element	ChemicalSubstance	ChemicalSubstance
/medicine/biofluid		Biofluid
/medicine/diagnostic_test		DiagnosticTest
/medicine/medical_device		MedicalDevice
/medicine/medical_specialty		MedicalSpecialty
	Medicine	Medicine
/education/educational_institution_campus /education/university	University	University
/organization/organization	Organisation	Organization
/education/field_of_study		FieldOfStudy
/people/ethnicity	EthnicGroup	Ethnicity
/sports	Sport	Sport
/time/holiday	Holiday	Holiday

Table 6.1: Translation of the types associated with entities from TextRazor in a unified version, for the domain of rare diseases

Since an entity could be tagged with multiple unified types, it is useful to consider the presence of a macroclass for display purposes.

6.5.2 Preprocessing

Italian language issues

Many NLP tasks are language dependent. Often the language limits the use of certain techniques or libraries, forcing the translation or the search for alternative solutions. So, the complexity of a single problem further increases when considering the need to deal with it in multilingual contexts. Only in recent years (or months sometimes) researchers have started working on extending the linguistic support of multiple models and solutions, thus increasing their applicability. The Italian language, which characterizes this case study, is poorly supported by online tools (mainly designed for English).

Social messages issues

Facebook text is known for its large, unbounded lexicon in addition to other distortions like misspellings, slangs and made up words. In social contexts, the concepts observed by patients are often described in imprecise way. Emojis are frequently used and can play a decisive role in interpreting the meaning of the text. Moreover, conversational messages are typically short. Understanding the quality characteristics of the text contained within the dataset is very important. In particular, it must be taken into account during the preprocessing transformations, since it could be decisive for the success of the analysis.

Quality preprocessing

To increase the quality of the data, the same pipeline defined in Section 4.3.1 was adopted, with the exception of the spelling correction and language translation phases. The latter constitutes a particularly complex task within the case study in question, where there are many particular terms, some vertical on the domain and others easily ambiguous (linked to names of doctors, specialized centers, etc.). For example, "policlinico gemelli" would be incorrectly translated into "twins polyclinic". Similarly, "dottor familiari" would be translated into "family doctors". Such a transformation would compromise the ability to interpret the result of the descriptive analysis. After several attempts, it was decided to limit the translation to the final sequence of terms related to the explanation of the phenomenon.

Analysis preprocessing

In this phase (immediately prior to the construction of the term-documents matrix) the following transformations took place: lower casing, removal of punctuation and numbers (with the exception of entity tags), and removal of extra white-spaces. A special mention must be made for stopwords removal. Given the variety of stoplists existing on the net for the Italian language, it was decided to merge them together in order to create a list as comprehensive as possible. In particular, those of `tm`, `quanteda`, `iso`[14], and the translated version of Hu and Liu have joined, together with the content of a custom file. The lemmatization was implemented but not used during the analysis for problems similar to those mentioned for the translation, as well as for the expensive computational time required in this case.

Documents classification

A common need among patients is to know whether other users' thoughts on a certain topic are positive or negative, and why. As verified with the survey (Section 6.2), the most common questions from this point of view concern medical treatments. In this case, therefore, the classification corresponds to an opinion mining task on the documents, and the description aims to explain the reasons why patients speak well or badly of a certain concept (local analysis) or of the disease in general (global analysis). As stated in [142], *"knowing that some documents have positive or negative opinions but not about what, is of limited use"*.

To estimate the opinion score associated with each document, we made use of a very simple algorithm based on **opinion words count**. Opinion words are words known for their semantic expression of polarity, commonly used to communicate positive or negative sentiments [142]. For example, "good", "excellent" and "amazing" are positive opinion words, and "bad", "poor", "terrible" are negative opinion words. In particular, we used the opinion lexicon (i.e., list of opinion words) published by Hu and Liu[15] (containing 6800 positive and negative words), appropriately translated into Italian to cope with the mismatch of the language. The specificity of the case study also made it appropriate to insert additional opinion words (both positive and negative).

```r
# Adds some other specific opinion words
# to positive and negative lists
pos.words = c(hu.liu.pos, 'rinata', 'ripresa', 'scende',
        'scendere', 'commosso', 'commossa', 'sorpreso', 'record',
```

[14] https://github.com/stopwords-iso/stopwords-iso

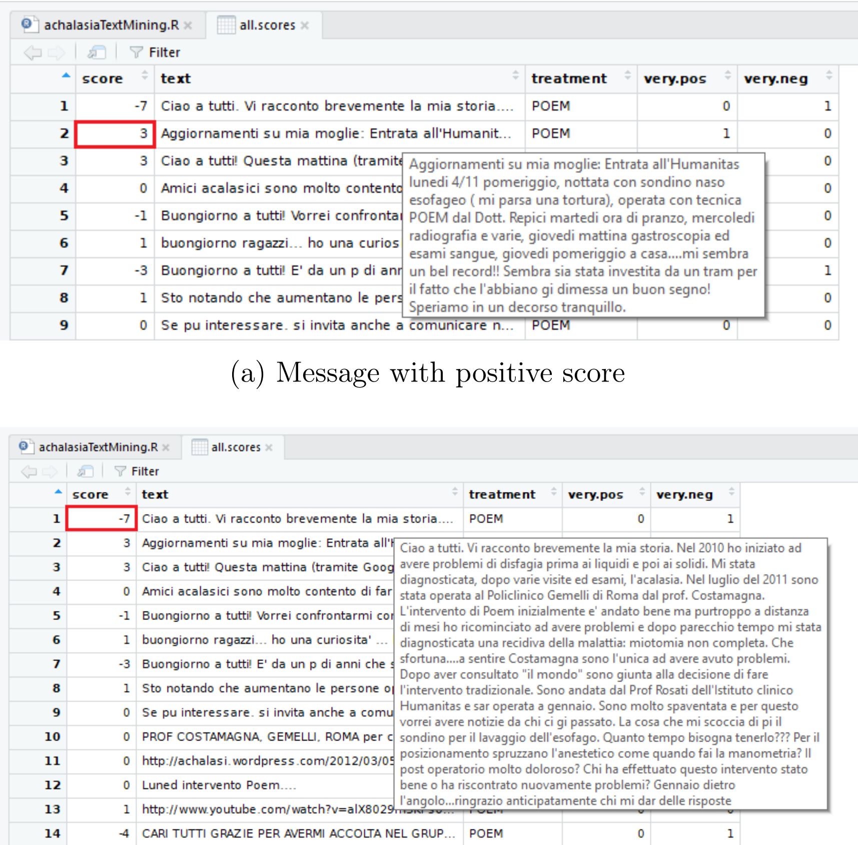
[15] https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html#lexicon

```
        'dimessa', 'dimesso', 'buon', 'buon segno', 'buona',
        'giovamento', ...)
neg.words = c(hu.liu.neg, 'recidiva', 'reflusso', 'patema',
        'bruciore', 'collaterale', 'spiazzante', 'acido',
        'sofferenza', 'tosse', 'singhiozzo', 'rigurgito', 'blocco',
        'fastidio', ...)
```

The score associated with each post and comment is therefore obtained by making the difference between the number of positive matches and the number of negative ones, as indicated by the following formula.

$$score(d) = nMathces(d, pos\_words) - nMatches(d, neg\_words) \tag{6.1}$$

Figure 6.7 gives a practical example related to the assignment of the opinion score.



	score	text	treatment	very.pos	very.neg
1	-7	Ciao a tutti. Vi racconto brevemente la mia storia....	POEM	0	1
2	3	Aggiornamenti su mia moglie: Entrata all'Humanit...	POEM	1	0
3	3	Ciao a tutti! Questa mattina (tramite			0
4	0	Amici acalasici sono molto contento			0
5	-1	Buongiorno a tutti! Vorrei confronta			0
6	1	buongiorno ragazzi... ho una curios			0
7	-3	Buongiorno a tutti! E' da un p di ann			1
8	1	Sto notando che aumentano le pers			0
9	0	Se pu interessare. si invita anche a comunicare n...	POEM	0	0

Aggiornamenti su mia moglie: Entrata all'Humanitas lunedi 4/11 pomeriggio, nottata con sondino naso esofageo (mi parsa una tortura), operata con tecnica POEM dal Dott. Repici martedi ora di pranzo, mercoledi radiografia e varie, giovedi mattina gastroscopia ed esami sangue, giovedi pomeriggio a casa....mi sembra un bel record!! Sembra sia stata investita da un tram per il fatto che l'abbiano gi dimessa un buon segno! Speriamo in un decorso tranquillo.

(a) Message with positive score

	score	text	treatment	very.pos	very.neg
1	-7	Ciao a tutti. Vi racconto brevemente la mia storia....	POEM	0	1
2	3	Aggiornamenti su mia moglie: Entrata all'H			
3	3	Ciao a tutti! Questa mattina (tramite Goog			
4	0	Amici acalasici sono molto contento di far			
5	-1	Buongiorno a tutti! Vorrei confrontarmi co			
6	1	buongiorno ragazzi... ho una curiosita' ...			
7	-3	Buongiorno a tutti! E' da un p di anni che s			
8	1	Sto notando che aumentano le persone o			
9	0	Se pu interessare. si invita anche a comu			
10	0	PROF COSTAMAGNA, GEMELLI, ROMA per c			
11	0	http://achalasi.wordpress.com/2012/03/05			
12	0	Luned intervento Poem....			
13	1	http://www.youtube.com/watch?v=alX8029			
14	-4	CARI TUTTI GRAZIE PER AVERMI ACCOLTA NEL GRUP...	POEM	0	1

Ciao a tutti. Vi racconto brevemente la mia storia. Nel 2010 ho iniziato ad avere problemi di disfagia prima ai liquidi e poi ai solidi. Mi stata diagnosticata, dopo varie visite ed esami, l'acalasia. Nel luglio del 2011 sono stata operata al Policlinico Gemelli di Roma dal prof. Costamagna. L'intervento di Poem inizialmente e' andato bene ma purtroppo a distanza di mesi ho ricominciato ad avere problemi e dopo parecchio tempo mi stata diagnosticata una recidiva della malattia: miotomia non completa. Che sfortuna....a sentire Costamagna sono l'unica ad avere avuto problemi. Dopo aver consultato "il mondo" sono giunta alla decisione di fare l'intervento tradizionale. Sono andata dal Prof Rosati dell'Istituto clinico Humanitas e sar operata a gennaio. Sono molto spaventata e per questo vorrei avere notizie da chi ci gi passato. La cosa che mi scoccia di pi il sondino per il lavaggio dell'esofago. Quanto tempo bisogna tenerlo??? Per il posizionamento spruzzano l'anestetico come quando fai la manometria? Il post operatorio molto doloroso? Chi ha effettuato questo intervento stato bene o ha riscontrato nuovamente problemi? Gennaio dietro l'angolo...ringrazio anticipatamente chi mi dar delle risposte

(b) Message with negative score

Figure 6.7: Examples of posts and opinion score assignment based on opinion words

Term Weighting

To highlight the importance of each term within the term-document matrix, it was decided not to adopt the classic tf-idf as weigthing scheme. In fact, a variation has been used (making use of **entropy**, as defined by Shannon) in the non-local part of the formula. Although apparently similar to idf, entropy leads to better results.

As stated in [235, 236], *"entropy global weighting is generally superior to normalized weighting, and both are better than the inverse document frequency function. The choice of the global weighting function affects the correlations more than any other characteristic. The use of idf global weighting produces correlations with human pairwise judgments that are uniformly worse than those achieved using entropy."*

```
tdmle <- lw_logtf(tdm) * (1 - entropy(tdm))
```

In our case we combine two factors:

- the term frequency log (i.e., occurrence log + 1);
- a factor inverse to the entropy of the term: high (i.e., close to 1) if the term appears in many documents and therefore is not very informative.

6.6 R packages and Python libraries

Below are the main R packages and Python libraries used for the implementation of the proposed text mining method (Table 6.2 and 6.3).

R package	Use within the project
`reticulate`	R and Python integration
plyr dplyr	Tools for efficiently manipulating datasets
`stringr`	Tools for string manipulation (e.g., replacement of occurrences of a matched pattern)
`qdap`	Multiple slang substitution
`purrr`	Functional programming toolkit
`ggplot2`	Plotting charts
`ggpubr`	Multiple charts combining
`doBy`	Data frame sorting
`tm`	Text Mining utilities (e.g., content transformation, common feature selection)
`quanteda`	Quantitative text analysis

`koRpus`	Wraps third party products, like TreeTagger for lemmatization and POS tagging
`lsa`	LSA analysis
`Rtsne`	t-SNE
`arrangements`	Iteration on permutations of list items
`lubridate`	Date filtering
`proxy`	Similarity / dissimilarity measures for ’dist()’
`dendextend`	Dendrogram editing
`pvclust`	Hierarchical clustering eith bootstrapped p-values
`reshape2`	Correlation matrix reshaping and heatmap creation
`tinytex`	LaTex installation
`Rserve`	Accessing R from Java

Table 6.2: Main R packages used for the implementation

Python library	**Use within the project**
`langdetect`	Language detection
`collections`	Most common item detection
`treetaggerwrapper`	Lemmatization
`symspellpy`	Spelling correction
`textblob`	Language translation
`re`	Regex handling

Table 6.3: Main Python libraries used for the implementation

6.7 Experiments

This section presents the results obtained through the application of the proposed Text Mining method on social data shared within the Facebook group *Acalasia esofagea... I malati “rari” non sono soli...!*.

It shows the extraction of knowledge through a global analysis (i.e., considering all posts and comments as a whole). The output of each step is examined and commented. The scientifically known correlations are thus researched within the latent semantic space to ascertain the effectiveness of the method.

6.7.1 Global Analysis

The R command to launch the analysis is as follows.

```
analysis_res <- automaticOpinionMiningLSA(
  documents = amae_data,
  entities = amae_ner_data,
  concept = list(
    concept_name = "all",
    concept_regex = "*"),
  min_creation_date = as.Date("2009-01-01"),
  max_creation_date = as.Date("2019-12-31"),
  lemmatization = FALSE,
  knee_point_calibration_mode = "lsa",
  known_terms_correlations = gold_standard_partition,
  starting_pos_query = NULL,
  starting_neg_query = NULL,
  qualityPreprocessingFun =
    function(texts) { return(cleanAndTranslate(texts)) },
  stoplist = complete_stopwords,
  opinionMiningFun = function(texts) { return(
    score.sentiment(
        texts,
        achalasia.pos.words,
        achalasia.neg.words,
        F, .progress = 'text')) },
  pos_opinion_min = 2,
  neg_opinion_max = -2,
  min_standard_term_freq = 0.001,
  min_entity_term_freq = 0.001,
  n_most_freq_terms = 20,
  n_concept_neighbors = 40,
  norm_threshold = 0.8,
  min_similarity_closeness = 0.9,
  semantically_closest_terms_n = 10,
  min_pvalue_threshold = 0.005)
```

As can be seen, there are no concept filters and the time range covers that of all data.

Number and distribution of documents

The total number of documents, considering the union of posts and comments, is 67,653. Their distribution over years is represented in Figure 6.8. The usage of the group has increased over time, reaching a peak in 2014. The number of shared posts and comments has then decreased again and has remained stable since 2016.

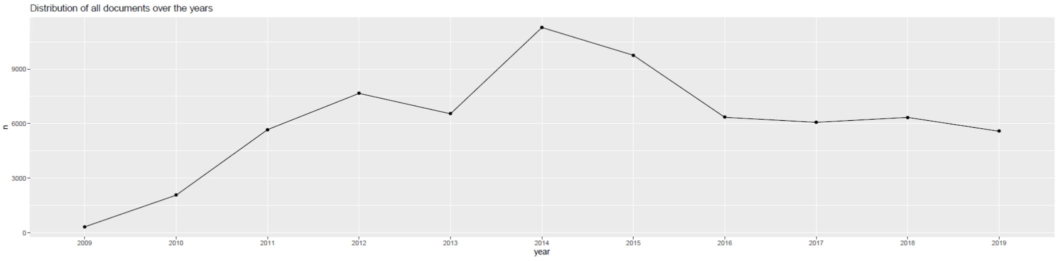
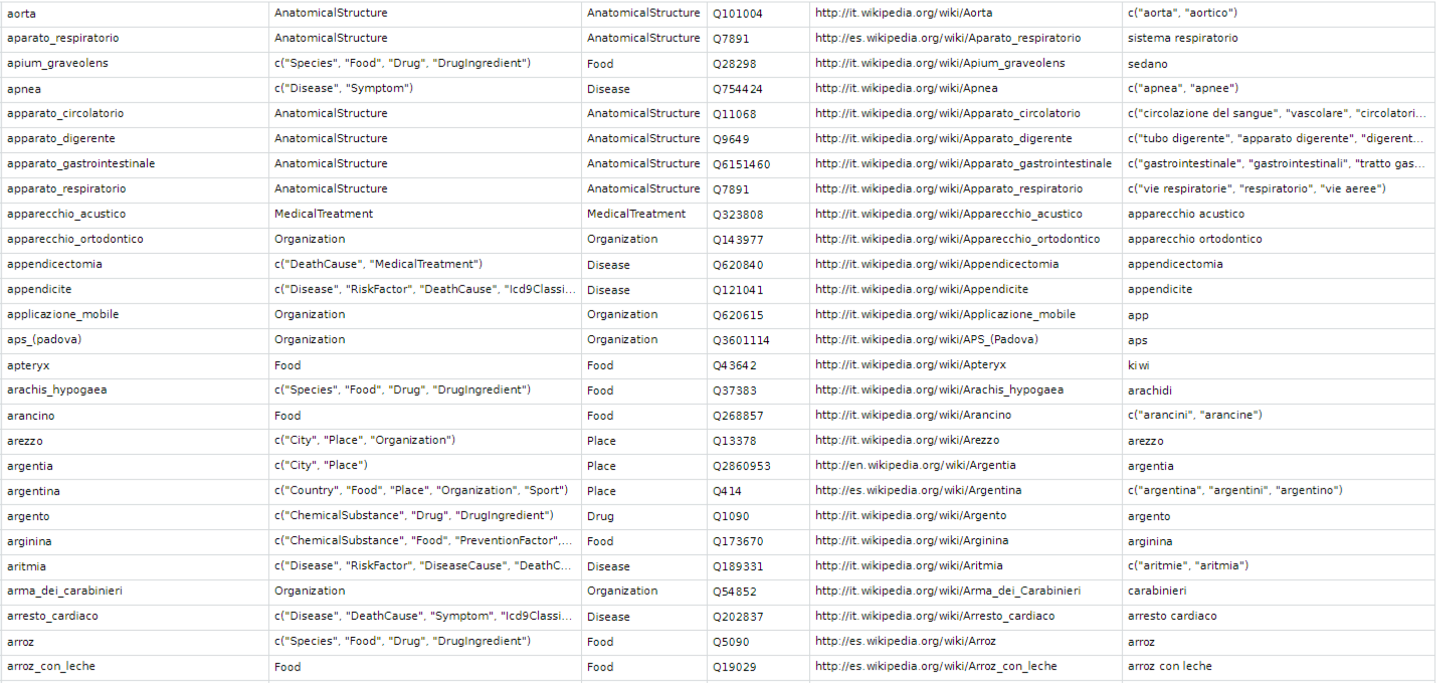


Figure 6.8: Global documents distribution over years

NER results before and after reconciliation

The total number of entities recognized by TextRazor's Entity Recognition in all documents is 41,958 (2,605 excluding duplicates). After the stages of reconciliation (the result of which is shown in Figure 6.9) and labeling of the text, the number of recognized and labeled entities is raised to 199,934.



aorta	AnatomicalStructure	AnatomicalStructure	Q101004	http://it.wikipedia.org/wiki/Aorta	c("aorta", "aortico")
aparato_respiratorio	AnatomicalStructure	AnatomicalStructure	Q7891	http://es.wikipedia.org/wiki/Aparato_respiratorio	sistema respiratorio
apium_graveolens	c("Species", "Food", "Drug", "DrugIngredient")	Food	Q28298	http://it.wikipedia.org/wiki/Apium_graveolens	sedano
apnea	c("Disease", "Symptom")	Disease	Q754424	http://it.wikipedia.org/wiki/Apnea	c("apnea", "apnee")
apparato_circolatorio	AnatomicalStructure	AnatomicalStructure	Q11068	http://it.wikipedia.org/wiki/Apparato_circolatorio	c("circolazione del sangue", "vascolare", "circolatori...
apparato_digerente	AnatomicalStructure	AnatomicalStructure	Q9649	http://it.wikipedia.org/wiki/Apparato_digerente	c("tubo digerente", "apparato digerente", "digerent...
apparato_gastrointestinale	AnatomicalStructure	AnatomicalStructure	Q6151460	http://it.wikipedia.org/wiki/Apparato_gastrointestinale	c("gastrointestinale", "gastrointestinali", "tratto gas...
apparato_respiratorio	AnatomicalStructure	AnatomicalStructure	Q7891	http://it.wikipedia.org/wiki/Apparato_respiratorio	c("vie respiratorie", "respiratorio", "vie aeree")
apparecchio_acustico	MedicalTreatment	MedicalTreatment	Q323808	http://it.wikipedia.org/wiki/Apparecchio_acustico	apparecchio acustico
apparecchio_ortodontico	Organization	Organization	Q143977	http://it.wikipedia.org/wiki/Apparecchio_ortodontico	apparecchio ortodontico
appendicectomia	c("DeathCause", "MedicalTreatment")	Disease	Q620840	http://it.wikipedia.org/wiki/Appendicectomia	appendicectomia
appendicite	c("Disease", "RiskFactor", "DeathCause", "Icd9Classi...	Disease	Q121041	http://it.wikipedia.org/wiki/Appendicite	appendicite
applicazione_mobile	Organization	Organization	Q620615	http://it.wikipedia.org/wiki/Applicazione_mobile	app
aps_(padova)	Organization	Organization	Q3601114	http://it.wikipedia.org/wiki/APS_(Padova)	aps
apteryx	Food	Food	Q43642	http://it.wikipedia.org/wiki/Apteryx	kiwi
arachis_hypogaea	c("Species", "Food", "Drug", "DrugIngredient")	Food	Q37383	http://it.wikipedia.org/wiki/Arachis_hypogaea	arachidi
arancino	Food	Food	Q268857	http://it.wikipedia.org/wiki/Arancino	c("arancini", "arancine")
arezzo	c("City", "Place", "Organization")	Place	Q13378	http://it.wikipedia.org/wiki/Arezzo	arezzo
argentia	c("City", "Place")	Place	Q2860953	http://en.wikipedia.org/wiki/Argentia	argentia
argentina	c("Country", "Food", "Place", "Organization", "Sport")	Place	Q414	http://es.wikipedia.org/wiki/Argentina	c("argentina", "argentini", "argentino")
argento	c("ChemicalSubstance", "Drug", "DrugIngredient")	Drug	Q1090	http://it.wikipedia.org/wiki/Argento	argento
arginina	c("ChemicalSubstance", "Food", "PreventionFactor",...	Food	Q173670	http://it.wikipedia.org/wiki/Arginina	arginina
aritmia	c("Disease", "RiskFactor", "DiseaseCause", "DeathC...	Disease	Q189331	http://it.wikipedia.org/wiki/Aritmia	c("aritmie", "aritmia")
arma_dei_carabinieri	Organization	Organization	Q54852	http://it.wikipedia.org/wiki/Arma_dei_Carabinieri	carabinieri
arresto_cardiaco	c("Disease", "DeathCause", "Symptom", "Icd9Classi...	Disease	Q202837	http://it.wikipedia.org/wiki/Arresto_cardiaco	arresto cardiaco
arroz	c("Species", "Food", "Drug", "DrugIngredient")	Food	Q5090	http://es.wikipedia.org/wiki/Arroz	arroz
arroz_con_leche	Food	Food	Q19029	http://es.wikipedia.org/wiki/Arroz_con_leche	arroz con leche

Figure 6.9: Example of entities data after reconciliation. For each entity there is the identifier corresponding to the local language, the list of unified types, the macro class, the Wikidata identifier, the link to Wikipedia, and the set of matched text that allowed recognition.

Documents classification

With the completion of the opinion mining task they are obtained 11,760 positive documents (score > 2) and 4,573 negative documents (score < -2), 72% and 28% respectively. The distribution of the opinions expressed in the documents, for each year, are shown in Figure 6.10. It can be observed that recently the positive posts have decreased significantly in proportion.

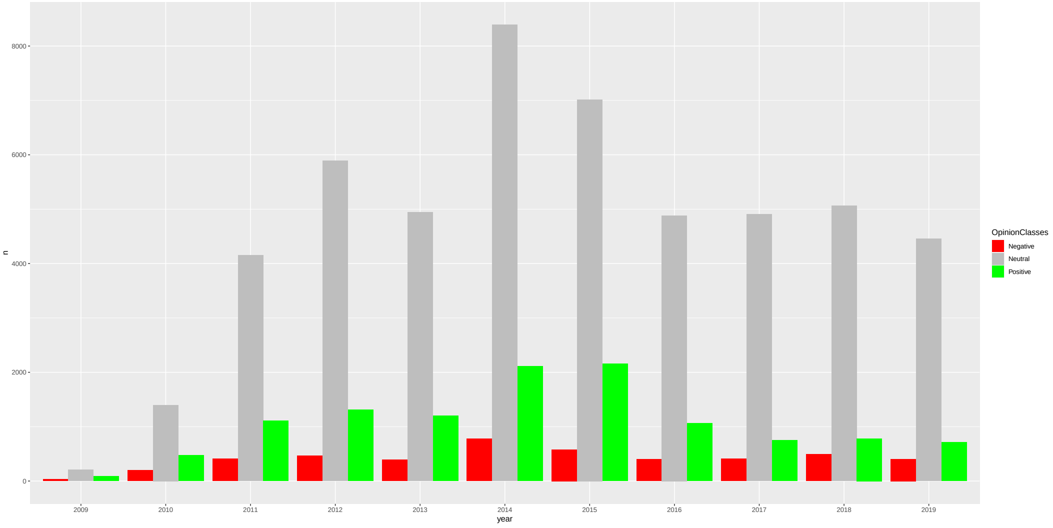


Figure 6.10: Global opinion distribution over years

Feature Selection

By applying the filter on the minimum frequency of terms (0.001), we obtain a document-term matrix made up of 1321 standard terms and only 192 entity terms.

Word Clouds

Frequency analysis, despite its simplicity, highlights several aspects of interest.

Figure 6.11 indicates how users frequently talk about treatments and diagnostic tests. From this point of view, manometry proves to be among the most discussed. In terms of treatments, pneumatic dilation has a higher number of occurrences than actual surgical interventions such as POEM and Heller Dor (despite the fact that the most cited doctor is Constantini, from Padua). Water, in its important role for the descent of food, is confirmed among the most mentioned terms.

Figure 6.12 is also interesting. The topic of treatments, requested by patients, is once again further confirmed in terms of importance on the data. Many more words with negative connotations appear from entities (such as pain,

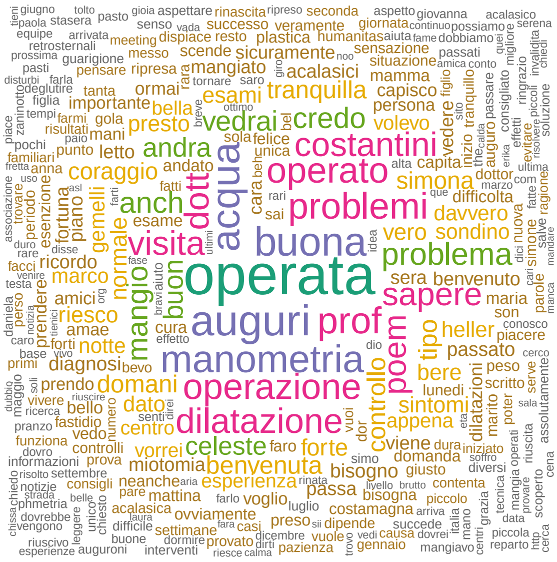


Figure 6.11: Word Cloud formed by standard terms (global analysis).

fear and gastroesophageal reflux). Padua (main reference center for Heller-Dor) is significantly more widespread than Rome (main reference center for POEM).

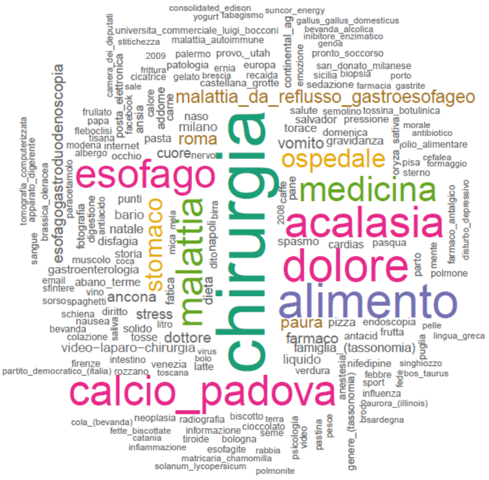


Figure 6.12: Word Cloud formed by entity terms (global analysis)

Figure 6.13 shows the Word Cloud given by the union of the two just analyzed (i.e., taking into account both standard and entities terms).

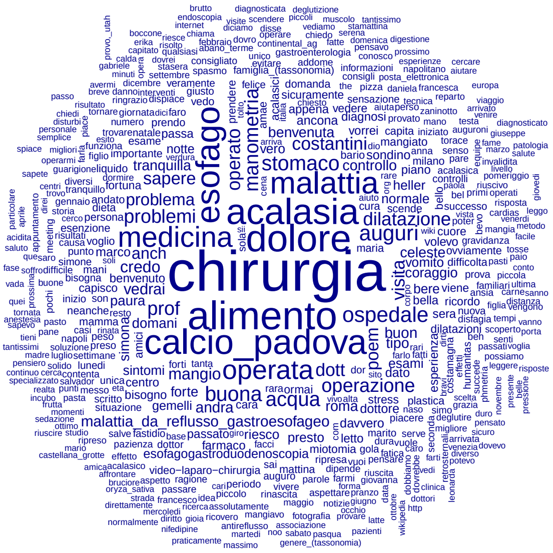


Figure 6.13: Word Cloud made up of both standard and entity terms (global analysis)

From Figure 6.14 it emerges that pain related to the disease is the most recurring term in documents labeled with a negative opinion class. The other symptoms of achalasia, such as regurgitation, reflux, chest pain, cough, muscle spasms, hiccups, headache and nausea (accompanied by stress, anxiety and fear) also find ample space. The ability to isolate such a large number of symptoms with frequency counting is undoubtedly relevant (without considering precision, observing the overall results). The most frequent terms in positive documents largely refer to messages of good wishes for a speedy recovery. Even in this case, however, Padua appears with higher frequency than Rome. This information also confirms how much the opinion mining classification based on word lexicon, although based on exact lexical matching and not capable of handling negations and propagations, can still lead to satisfactory results.

LSA Execution

Within the experiment we chose to use LSA as Language Model. For timing reasons, the calibration by Gold Standards was adopted only for the choice of the knee point (and not also for the determination of the number of dimensions k inside the new space). Consequently, the heuristic released within the `lsa` package was used to choose k, applying some modifications with repeated tests aimed at identifying a suitable value for the analysis.

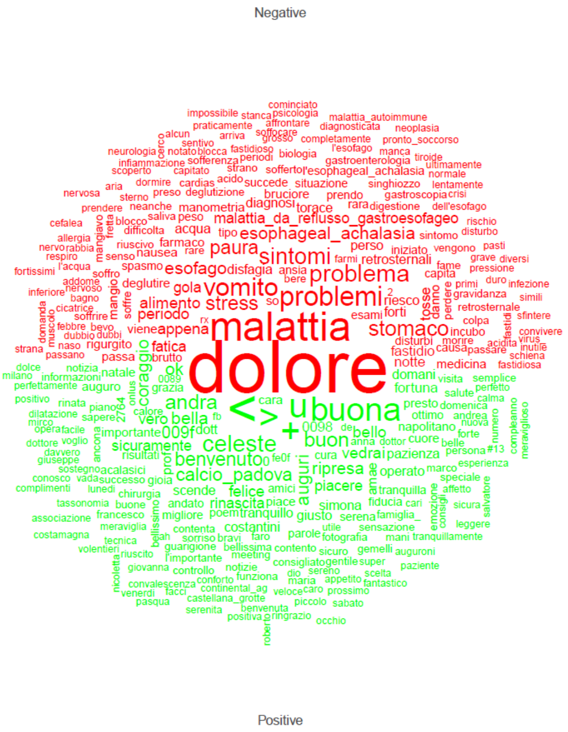


Figure 6.14: Word Cloud of comparison between the terms present in positive and negative documents (global analysis)

We used $k = 100$.
Automatic calibration selects 5 as the minimum capable of capturing the best known correlations expressed within the Gold Standards, similarly to an interactive approach (Figure 6.15). 2D visualizations are reported in Figures 6.16 and 6.17.

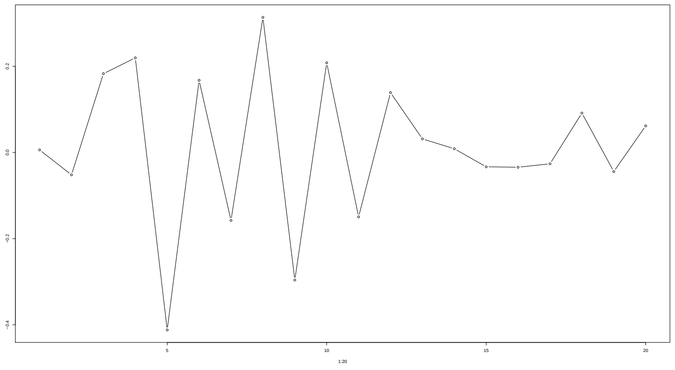


Figure 6.15: Curvature function for global analysis

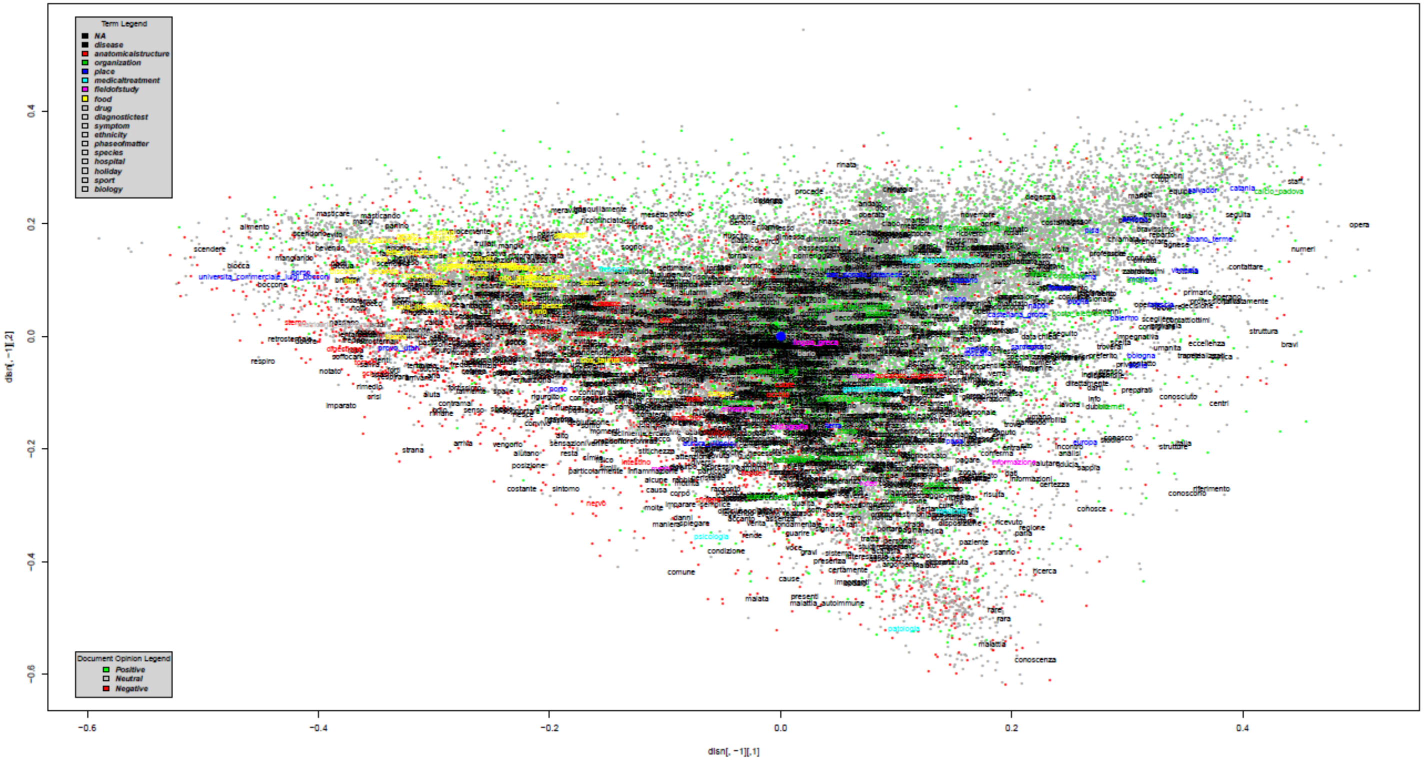


Figure 6.16: Global Analysis, LSA on dimensions 2 & 3

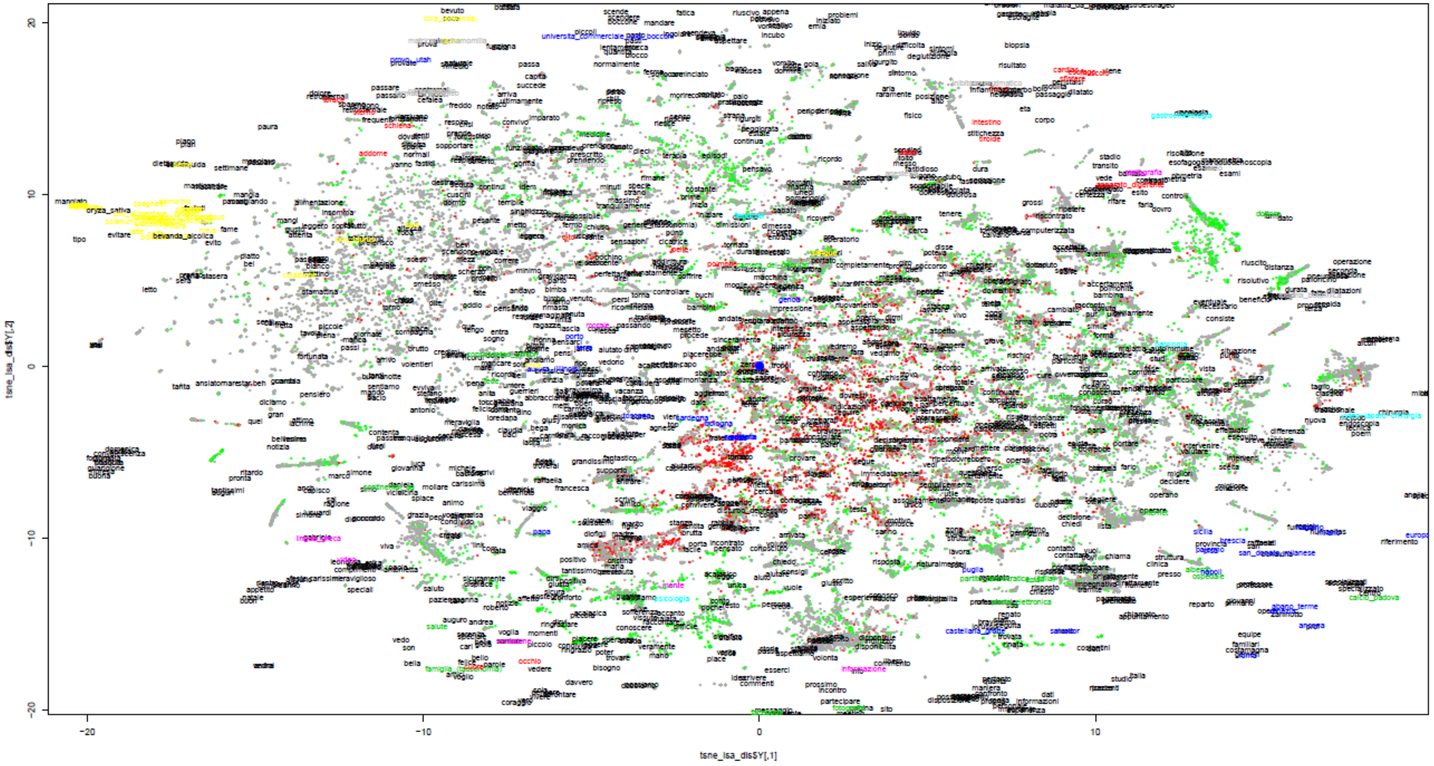
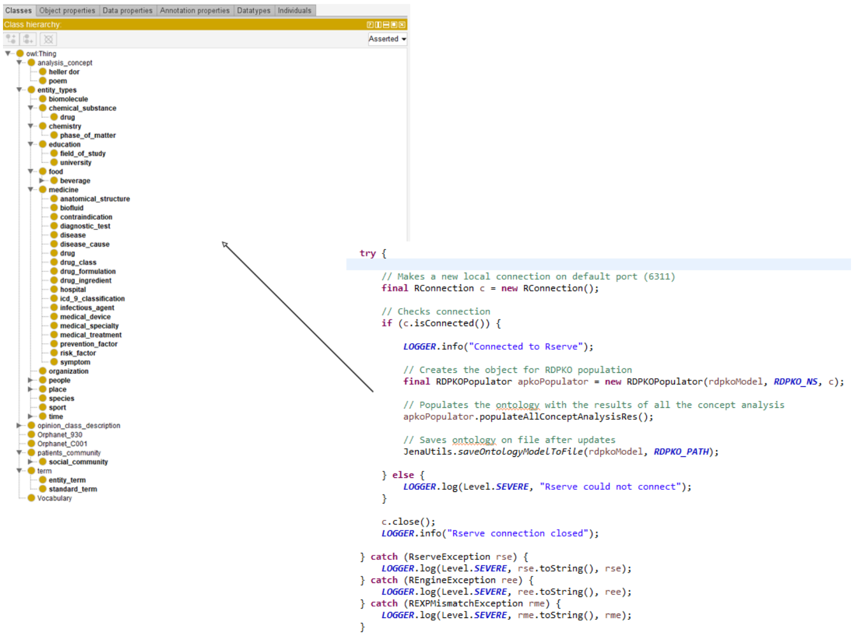


Figure 6.17: Global Analysis, t-SNE with perplexity = 20

As can be seen from the two-dimensional graphic representation, within the latent semantic space two areas of concentration can be recognized, distinct for the positive (right) and negative (left) classes of opinion. Continuing with the automatic analysis, we can see how the terms characterizing the negative area concern the difficulty in descending the food and the consequent retrosternal pains. Conversely, those who most identify the positive concentration refer to the doctors of Padua and Rome with their respective intervention techniques.

6.8 Knowledge Graph Construction

For the realization of the knowledge graph we have chosen to work with Protégé[16] and Apache Jena[17]. Through the latter framework, it was possible to connect to the R session with the aim to submit the commands necessary to launch the analysis (both globally and locally). Communication between Java and R was made possible by RServe[18]. It acts as a socket server (TCP/IP or local sockets) which allows binary requests to be sent to R.



```
owl:Thing
    analysis_concept
        heller dor
        poem
    entity_types
        biomolecule
        chemical_substance
            drug
        chemistry
            phase_of_matter
        education
            field_of_study
            university
        food
            beverage
        medicine
            anatomical_structure
            biofluid
            contraindication
            diagnostic_test
            disease
            disease_cause
            drug
            drug_class
            drug_formulation
            drug_ingredient
            hospital
            icd_9_classification
            infectious_agent
            medical_device
            medical_specialty
            medical_treatment
            prevention_factor
            risk_factor
            symptom
        organization
        people
        place
        species
        sport
        time
    opinion_class_description
    Orphanet_930
    Orphanet_C001
    patients_community
        social_community
    term
        entity_term
        standard_term
    Vocabulary
```

```java
try {

    // Makes a new local connection on default port (6311)
    final RConnection c = new RConnection();

    // Checks connection
    if (c.isConnected()) {

        LOGGER.info("Connected to Rserve");

        // Creates the object for RDPKO population
        final RDPKOPopulator apkoPopulator = new RDPKOPopulator(rdpkoModel, RDPKO_NS, c);

        // Populates the ontology with the results of all the concept analysis
        apkoPopulator.populateAllConceptAnalysisRes();

        // Saves ontology on file after updates
        JenaUtils.saveOntologyModelToFile(rdpkoModel, RDPKO_PATH);

    } else {
        LOGGER.log(Level.SEVERE, "Rserve could not connect");
    }

    c.close();
    LOGGER.info("Rserve connection closed");

} catch (RserveException rse) {
    LOGGER.log(Level.SEVERE, rse.toString(), rse);
} catch (REngineException ree) {
    LOGGER.log(Level.SEVERE, ree.toString(), ree);
} catch (REXPMismatchException rme) {
    LOGGER.log(Level.SEVERE, rme.toString(), rme);
}
```

Figure 6.18: From Apache Jena to Protégé through RServe

[16] https://protege.stanford.edu/
[17] https://jena.apache.org/
[18] https://cran.r-project.org/web/packages/Rserve/index.html

Conclusions and Future Challenges

6.9 Discussion and Conclusions

This thesis has presented two main contributions, respectively in the research areas of Text Mining and Semantic Web. In the first one, we have proposed a novel unsupervised method of descriptive analysis. By modeling terms and documents together in a latent semantic space, we have shown how it is possible to automatically generate textual explanations for a phenomenon of interest, starting from unlabeled data and accompanying the results with accurate probabilistic information. Secondly, we dealt with the problem linked to the representation of knowledge extracted with the first contribution, demonstrating how the flexibility and modularity of the latter can be exploited to define a knowledge graph learning methodology.

We demonstrated the effectiveness of the work through a case study in the medical field, specifically focused on Esophageal Achalasia. We conducted experiments on short conversational posts mainly shared by patients and caregivers, using LSA as Language Model. Through the discovery of statistically significant evidences, the analyzes have allowed the identification of scientific medical correlations directly from the patient's posts. Finally, it was possible to create a knowledge graph, extending the disease node on ORDO and adopting a patient-centered vision.

As future works, we plan to apply the descriptive text mining method to Language Models based on neural networks (comparing the results with the current ones). Moreover, we are going to develop a Shiny[19] application to make available the work done. We also intend to extend the work on the ontological part, introducing tags towards HPO and HOOP. Research and experiments on the application of contributions to subword tokenization (and not at word level) could also significantly improve the ability to highlight the presence of latent semantic relationships between data. Finally, the thesis also lends itself to an extension towards BigData technologies, with the aim not only of creating a knowledge graph automatically, but also of maintaining and evolving

[19] `https://shiny.rstudio.com/`

it dynamically over time.

6.10 Open Challenges and Future Potential

It is also observed that the issue of explainability treated in this thesis is extremely topical. The trend of research towards "breaking the black box" has also recently been expressed by Yoav Goldberg (Figure 6.19). He points out that, in the evolution of NLP, solutions have moved from rule-based systems (patterns) to models of Machine Learning (ML) and Deep Learning (DL), with a loss of transparency and with a substantial increase in labeled data required. In the next few years the goal is to have tools that do not require users to be experts in linguistics or automatic learning. One way to achieve this is to work on returning to rule-based systems, this time aided by ML / DL with the hope of obtaining models that are again transparent, understandable, debuggable, easy to control, fast and requiring little labeled data.

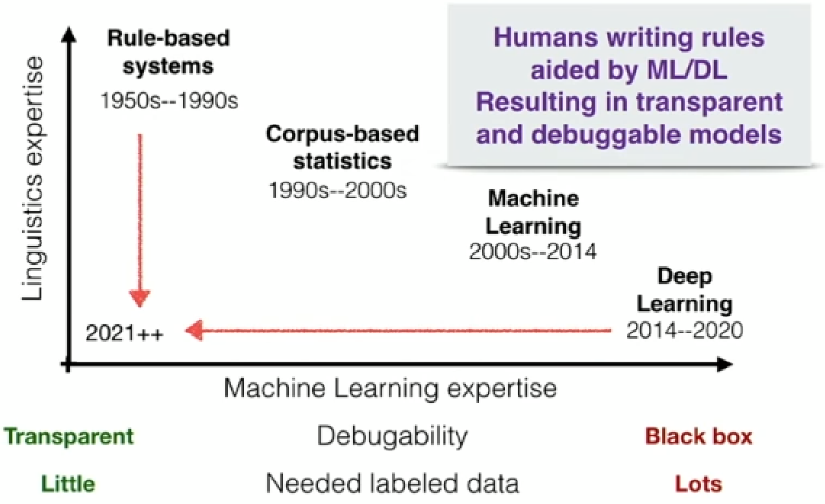


Figure 6.19: The importance of understanding what deep learning models are doing

Obtained from "The missing elements in NLP (spaCy IRL), Yoav Goldberg, July 2019"

Appendix A

Code examples

A.1 Latent Semantic Analysis Utils

```
#' Performs a normalization (to unit vector conversion) of each
#' row of a dataframe
#'
#' @param x The dataframe with unnormalized vectors as rows
#' @return The dataframe with normalized vectors
normRows <- function(x) x / apply(x, 1, norm, "2")


#' Calculates the cosine similarities between a vector and
#' each row of a matrix
#'
#' @param x The matrix
#' @param q The vector
#' @return The vector with the cosine similarities
cosines <- function(x, q) apply(x, 1, cosine, as.vector(q))


#' Gets the highest n values of a vector
#'
#' @param x The vector
#' @param n The number of top item to consider
#' @return The top n items inside x
top <- function(x, n) order(x, decreasing = TRUE)[1:n]


#' Gets the terms in the latent space with a norm higher
#' than a threshold
```

```
#'
#' @param tls_coordinates The coordinates of the terms in the
#' latent space
#' @param tls_names The names of the terms (textual representation)
#' @param norm_threshold The minimum norm for a relevant term
#' @return A complete dataframe with term-norm pairs (tls_norms)
#' and another one with the highest norm terms (tls_high_norms)
searchHighNormTerms <- function(tls_coordinates, tls_names,
    norm_threshold) {

  # Calculates the norm of the terms in the latent space
  tls_norms <- apply(tls_coordinates, 1, norm, "2")
  tls_norms <- data.frame(
    term = tls_names,
    norm = tls_norms,
    row.names = NULL)

  # Sorts decreasingly terms in the selected dimensions
  # of the latent space by their norm
  tls_norms <- orderBy(~-norm, tls_norms)

  # Filters only the terms with a sufficiently high norm
  tls_high_norms <- tls_norms %>% filter(norm > norm_threshold)

  list(tls_norms = tls_norms, tls_high_norms = tls_high_norms)

}


#' Performs all the preliminary operations for the calculation
#' of the similarity between a query and the terms in the
#' latent space
#'
#' @param q The query
#' @param tdm The term document matrix
#' @param lsa The components of the latent semantic space after SVD
#' @return
#' - The original string query (q)
#' - The binary query document in the latent space (bin_q)
#' - The query document after tf-idf entropy weighting (w_q)
#' - The query document in the latent space (ls_q)
#'   and its normalization (lsn_q)
#' - The query document equivalent to a V matrix row (dk)
#' - The V * Sigma^1/2 element for semantic similarity calculation
#'   between query and terms (dksrs)
makeQuery <- function(q, tdm, lsa) {

  # Creates the query vector (binary vector)
  bin_q <- lsa::query(q, rownames(tdm))

  # The query is like a new document to add to the latent space
  # So it applies all the transformations made to those already
  # inside (term weighting) in order to fold it into the LSA space
  w_q <- lw_logtf(bin_q) * (1 - entropy(tdm))

  # Calculates the position of the query in the latent space
  # (V * Sigma)
  ls_q <- t(w_q) %*% lsa$tk

  # Calculates the normalized query vector in latent space
  # for visualization purposes
  lsn_q <- normRows(ls_q)

  # Transforms the query vector in a new document (row of V matrix)
  # q_k = q^T * U_k * Sigma_k^-1
  dk <- ls_q %*% diag(lsa$sk ^ -1)

  # The similarity between a query and a terms is calculated as
  # cosine(V * Sigma^1/2, U * Sigma^1/2) =
  # cosine(dk %*% diag(sqrt(lsa$sk), lsa$tk %*% diag(sqrt(lsar$sk)))
  # dksrs is so one of the two necessary elements for similarity
  # calculation
  dksrs <- dk %*% diag(sqrt(lsa$sk))

  # Returns a named list with the results
  list(
    q = q, bin_q = bin_q, w_q = w_q,
    ls_q = ls_q, lsn_q = lsn_q,
    dk = dk, dksrs = dksrs)

}


#' Calculates the correlation between a query and a class
#' inside a latent space, using chi-squared test
#'
#' @param q The query to consider for correlation calculation
#' @param tdm The term document matrix
#' @param lsa The components of the latent semantic space after SVD
#' @param dls The documents' positions in the latent space
#' @param lsa_dimensions The LSA dimensions to use for the computation
#' @param class The class to consider for correlation
#' @param classes The classes related to each document of the entire
#' dataset
#' @return The chi-squared result for the correlation between
#' q and class, and the query object.
calculateQueryClassCorrelationLSA <- function(q, tdm, lsa, dls,
    lsa_dimensions, class, classes) {

  # Converts the query as a document in the LSA space (query document)
  qd <- makeQuery(q, tdm, lsa)

  # Calculates the number of occurrences of the target class
  # between all classes. It fixes the maximum number of results
  # returned from the ranking-based semantic search model (R precision)
  class_count <- sum(classes == class)

  # Verifies objectively if there is a semantic correlation between
  # the query and the class. In the latent space, the documents
  # considered could also not contain the terms inside the query
  # but terms semantically related to them

  # Builds the contingency table
  q.vs.class <- table(
    1:nrow(dls) %in%
        top(cosines(
            dls[, lsa_dimensions],
            qd$ls_q[lsa_dimensions]),
        class_count),
    classes == "Positive")
  dimnames(q.vs.class) <- list(
    q = c("No", "Si"), class = c("No", "Si"))
  print(as.character(q))
  print(q.vs.class)

  # Performs a chi-squared test between the query and the class
  chisqtest <- chisq.test(q.vs.class, correct = FALSE)
  cat('Expected: ', chisqtest$expected, "\n")
  cat('Observed: ', chisqtest$observed, "\n")

  list(chisqtest = chisqtest, query = qd)
```

```
}


#' Performs a descriptive analysis for a certain opinion class
#' in LSA space.
#'
#' @param starting_query Optional. The query document from which
#' start the analysis
#' @param opinion_class The opinion class for which build a
#' description
#' @param opinion_class_count The number of documents with the
#' opinion class to analyze
#' @param opinion_classes The opinion class of each document
#' @param docs_ids The identifier of each document
#' @param tls_norms The norm of each term in the latent space
#' @param tls_high_norms Terms with sufficiently high norm in the
#' latent space
#' @param lsa The components of the latent semantic space after SVD
#' @param lsa_dimensions The LSA dimensions to consider during the
#' analysis
#' @param dls The documents' positions in the latent space
#' (used for computation)
#' @param dlsn The documents' normalized positions in the latent space
#' (used for 2D representation)
#' @param tlsn The terms' normalized positions in the latent space
#' (used for 2D representation)
#' @param tksrs The U * Sigma^1/2 element for semantic similarity
#' calculation between query and terms
#' @param min_similarity_closeness The minimum value for cosine
#' similarity that must be satisfied during step-by-step analysis
#' @param semantically_closest_terms_n The number of semantically
#' closest terms to consider during step-by-step analysis
#' @param min_pvalue_threshold The minimum pvalue that must be
#' satisfied by a class description
#' @return The list with the results of the descriptive analysis:
#' - The description for the opinion class in original language (cd)
#'   and in english (cd_en)
#' - The p-value score related to the opinion class description
#'   (pvalue)
#' - The position of the opinion class description document in the
#'   latent space (ls)
#' - The normalized position of the opinion class description document
#'   in the latent space (lsn)
descriptiveAnalysisLSA <- function(
    starting_query = NULL,
    opinion_class, opinion_class_count,
    opinion_classes,
    docs_ids,
    tls_norms, tls_high_norms,
    tdm,
    lsa, lsa_dimensions,
    dls, dlsn, tlsn,
    tksrs,
    min_similarity_closeness,
    semantically_closest_terms_n,
    min_pvalue_threshold) {

  if (is.null(starting_query)) {

    # The starting query document is not specified...

    # For each term with an hign norm, performs a chi-squared test
    # in order to verify the correlation (low p-value) between
    # them and the opinion class
    tls_high_norms_chisqtest_class <- lapply(
      tls_high_norms$term,
      function(x) calculateQueryClassCorrelationLSA(
        x, tdm, lsa, dls, lsa_dimensions,
        opinion_class, opinion_classes))

    tls_high_norms_chisqtest_class <- data.frame(
      term = tls_high_norms$term,
      norm = tls_high_norms$norm,
      pvalue = unlist(
        lapply(
            lapply(
                tls_high_norms_chisqtest_class,
                `[[`, "chisqtest"),
            `[[`, "p.value"),
        use.names = FALSE),
      # The I marker allows the creation of a dataframe without
      # a new column for each coordinate but with only one
      # vector column
      ls = I(lapply(
        lapply(tls_high_norms_chisqtest_class, `[[`, "query"),
        `[[`, "ls_q")),
      lsn = I(lapply(
        lapply(tls_high_norms_chisqtest_class, `[[`, "query"),
        `[[`, "lsn_q")),
      dksrs = I(lapply(
        lapply(tls_high_norms_chisqtest_class, `[[`, "query"),
        `[[`, "dksrs"))
    )

    # Orders high norm terms from those most correlated
    # with the opinion_class to those less correlated
    tls_high_norms_chisqtest_class <- orderBy(
        ~pvalue, tls_high_norms_chisqtest_class)

    # Selects the high norm term with the lowest pvalue
    # as the starting one
    top_pvalue_term <- head(tls_high_norms_chisqtest_class, 1)
    top_pvalue_term <- list(
      text = as.character(top_pvalue_term$term[[1]]),
      pvalue = top_pvalue_term$pvalue[[1]],
      ls = top_pvalue_term$ls[[1]],
      lsn = top_pvalue_term$lsn[[1]],
      dksrs = top_pvalue_term$dksrs[[1]]
    )

    # Initializes opinion_class description query data
    # with top pvalue term data
    oc_description_query <- top_pvalue_term

  } else {

    # The starting query document is already specified...

    # Calculates the correlation between
    # the starting query document and the opinion class
    sq_chisqtest_opinion_class <- calculateQueryClassCorrelationLSA(
      starting_query, tdm, lsa, dls, lsa_dimensions,
      opinion_class, opinion_classes)
    sq <- list(
      text = starting_query,
      pvalue = (sq_chisqtest_opinion_class[["chisqtest"]])[["p.value"]],
      ls = (sq_chisqtest_opinion_class[["query"]])[["ls_q"]],
      lsn = (sq_chisqtest_opinion_class[["query"]])[["lsn_q"]],
      dksrs = (sq_chisqtest_opinion_class[["query"]])[["dksrs"]]
    )

    # Initializes the opinion_class description query
    oc_description_query <- sq

  }

  analysis_finished <- FALSE
  while (!analysis_finished) {

    # Checks how opinion classes are distributed in the
    # semantic search result compared to the current
    # opinion_class description, showing a graphical
    # distribution
    nearest <- sort(
      cosines(
        dls[, lsa_dimensions],
        oc_description_query$ls[lsa_dimensions]),
      decreasing = T)
    damcols2 <- rep("black", length(docs_ids))
    damcols2[opinion_classes == opinion_class] <- "red"
    plot(1:opinion_class_count, nearest[1:opinion_class_count],
         pch = 20, cex = 0.7,
         col = damcols2[match(names(nearest), docs_ids)])

    # Removes from U * Sigma^1/2 the previously selected terms
    # for opinion_class description in order to not select
    # them again
    tksrs_ncd <- tksrs[!(rownames(tksrs) %in%
        unlist(base::strsplit(oc_description_query$text, " "))), ]

    # Calculates the semantic similarity between the query and
    # the terms, and order them
    # cosine(V * Sigma^1/2, U * Sigma^1/2)
    # cosine between -1 and 1 because the vector space also
    # includes the negative quadrants
    closeness_terms_ranking <- sort(
      cosines(tksrs_ncd[, lsa_dimensions],
        oc_description_query$dksrs[lsa_dimensions]),
      decreasing = TRUE)

    # Selects the N semantically nearest terms that satisfy
    #a minimum closeness threshold
    nearest_terms <- names(
      head(
        closeness_terms_ranking[closeness_terms_ranking >
            min_similarity_closeness],
        semantically_closest_terms_n))

    # Sorts the nearest terms by norm length descending order
    nearest_terms_norms <- dplyr::filter(
        tls_norms,
        term %in% nearest_terms)
    nearest_terms_norm_ranking <- orderBy(~-norm, nearest_terms_norms)

    # Navigates the nearest terms from the one with the highest norm
    # to the one with the lowest norm. The first met term with a
    # sufficient low pvalue is used for the query extension
    # representing the opinion class description
    new_term_chisqtest <- NULL
    i <- 1
    while (is.null(new_term_chisqtest) && i <= length(nearest_terms)) {

      # Creates the new query temp (opinion_class description)
      new_pos_description <- paste(
        oc_description_query$text, as.character(nearest_terms[[i]]))

      # Calculates the correlation between the new opinion class
      # description and the opinion class
      near_term_chisqtest <- calculateQueryClassCorrelationLSA(
        new_pos_description, tdm, lsa, dls, lsa_dimensions,
        opinion_class, opinion_classes)

      # Gets the pvalue of the chisquared test
      near_term_pvalue <-
        (near_term_chisqtest[["chisqtest"]])[["p.value"]]

      # Checks if the pvalue is low enough
      if (near_term_pvalue < min_pvalue_threshold) {
        new_term_chisqtest <- near_term_chisqtest
      }

      # Goes to the next nearest term with a lower norm
      i <- i + 1

    }

    if (is.null(new_term_chisqtest)) {
      # If no one of the nearest terms (independently by their norm)
      # has a pvalue low enough, stops the descriptive analysis
      # (opinion class description completed)
      analysis_finished <- TRUE
    } else {
      # Updates opinion class data with the selected term
      oc_description_query <- list(
        text = (new_term_chisqtest[["query"]])[["q"]],
        pvalue = (new_term_chisqtest[["chisqtest"]])[["p.value"]],
        ls = (new_term_chisqtest[["query"]])[["ls_q"]],
        lsn = (new_term_chisqtest[["query"]])[["lsn_q"]],
        dksrs = (new_term_chisqtest[["query"]])[["dksrs"]]
      )

    }

  }

  list(
    cd = oc_description_query$text,
    cd_en = translate(oc_description_query$text, "en"),
    cd_en = oc_description_query$text,
    pvalue = oc_description_query$pvalue,
    ls = oc_description_query$ls,
    lsn = oc_description_query$lsn)

}
```

Appendix B

Gold Standards Sample

Thanks to Dr. Cristina Lanni (University of Pavia) for the collaboration in the realization of the Gold Standards with references, where possible, to scientific publications.

B.1 Disease - Place

1. **acalasia ↔ padova**
   Costantini M, Salvador R, Capovilla G, Vallese L, Costantini A, Nicoletti L, Briscolini D, Valmasoni M, Merigliano S. A Thousand and One Laparoscopic Heller Myotomies for Esophageal Achalasia: a 25-Year Experience at a Single Tertiary Center. J Gastrointest Surg. 2019 Jan;23(1):23-35. doi: 10.1007/s11605-018-3956-x. Epub 2018 Sep 20.

2. **acalasia ↔ roma**
   Familiari P, Gigante G, Marchese M, Boskoski I, Tringali A, Perri V, Costamagna G. Peroral Endoscopic Myotomy for Esophageal Achalasia: Outcomes of the First 100 Patients With Short-term Follow-up. Ann Surg. 2016 Jan;263(1):82-7. doi: 10.1097/SLA.0000000000000992.

3. **acalasia ↔ brescia**

4. **acalasia ↔ rozzano, milano**

5. **acalasia ↔ abano_terme**

6. **acalasia ↔ pisa**

7. **acalasia ↔ napoli**

8. **acalasia ↔ bologna**

B.2 Disease - Symptom

1. **acalasia ↔ disfagia**
   Clavé P, Shaker R. Dysphagia: current reality and scope of the problem. Nat Rev Gastroenterol Hepatol. 2015 May;12(5):259-70. doi: 10.1038/nrgastro.2015.49.
2. **acalasia ↔ disfagia, solidi**
3. **acalasia ↔ disfagia, liquidi**
4. **acalasia ↔ disfagia, notte**
5. **acalasia ↔ tosse**
6. **acalasia ↔ infezioni**
7. **acalasia ↔ calo, peso**
8. **acalasia ↔ angina, esofagea**
9. **acalasia ↔ dolore, retrosternale, torace**
10. **acalasia ↔ singhiozzo**
11. **acalasia ↔ eruttazione**
12. **acalasia ↔ rigurgito**
13. **acalasia ↔ vomito**
14. **acalasia ↔ blocco, bolo**
15. **acalasia ↔ peristalsi**
16. **acalasia ↔ anemia**
17. **acalasia ↔ alitosi**
18. **acalasia ↔ debolezza**
19. **acalasia ↔ inappetenza**
20. **acalasia ↔ fitte**

B.3 Disease - Doctor

1. **acalasia ↔ peracchia**
   `http://amsdottorato.unibo.it/8502/1/Tassi_Valentina_tesi.pdf`
2. **acalasia ↔ fumagalli**
   `https://www.humanitas.it/news/14253-in-humanitas-un-equipe-dedicata-contro-le-malattie-di-esofago-e-stomaco`

3. **acalasia ↔ merigliano**
   http://www.singem.it/wp-content/uploads/2018/02/Programma-_-Miotomia-Padova-2017.pdf
4. **acalasia ↔ penagini**
   https://work.unimi.it/chiedove/cv/roberto_penagini.pdf
5. **acalasia ↔ bonavina**
   http://www.spec-chir.it/wp-content/uploads/2014/12/BONAVINA-CV.pdf
6. **acalasia ↔ santi**
   https://core.ac.uk/download/pdf/79618984.pdf
7. **acalasia ↔ savarino**
   https://en.didattica.unipd.it/off/docente/DA2C61BE134B0CD18B9D936D750F1023
8. **acalasia ↔ sarnelli**
9. **acalasia ↔ ancona**
   https://www.ermannoancona.eu/patologie/acalasia/
10. **acalasia ↔ zaninotto**
11. **acalasia ↔ salvador**
12. **acalasia ↔ costantini**
13. **acalasia ↔ repici**
14. **acalasia ↔ costamagna**
15. **acalasia ↔ rosati**
    https://www.humanitasalute.it/prima-pagina-ed-eventi/63936-acalasia-al-via-uno-studio-multicentrico/
16. **acalasia ↔ capizzi**
    http://556662.rivera2019.com/195213-AKWUCJTESK-Acalasia-esofagea-Criticita-Daniele-Capizzi-Francesco-Domenico-Capizzi/
17. **acalasia ↔ mattioli** https://sandromattioli.it/Patologie/Acalasia

B.4 SpecializedCenter - Doctor

1. **humanitas, milano, rozzano ↔ peracchia**
2. **humanitas, milano, rozzano ↔ fumagalli**
3. **brescia ↔ fumagalli**

4. **padova** ↔ **ancona**
5. **padova** ↔ **merigliano**
6. **padova** ↔ **roul**
7. **padova** ↔ **costantini**
8. **padova** ↔ **salvador**
9. **padova** ↔ **savarino**
10. **padova** ↔ **zaninotto** `http://www.sied.it/clients/www.sied.it/public/files/Acalasiaesofageastrategieditrattamento.pdf`
11. **abano_terme** ↔ **ancona**
12. **abano_terme** ↔ **zaninotto**
13. **pisa** ↔ **santi**
14. **napoli** ↔ **sarnelli**
15. **milano** ↔ **penagini**
16. **gemelli** ↔ **costamagna**
17. **gemelli** ↔ **famigliari**

Bibliography

[1] Orphanet. The portal for rare diseases and orphan drugs. https://www.orpha.net. Accessed 4 Feb 2020.

[2] GlobalGenes. Rare diseases: facts and statistics. https://globalgenes.org/rare-facts/. Accessed 4 Feb 2020.

[3] Daina E. Aperia A. Schieppati A., Henter J.I. Why rare diseases are an important medical and social issue. *Lancet*, 371:2039–2041, June 2008.

[4] Kyte D. Pankhurst T. Kerecuk L. Ferguson J. Lipkin G. Calvert M. Slade A., Isa F. Patient reported outcome measures in rare diseases: a narrative review. *Orphanet Journal of Rare Diseases*, 13:61, April 2018.

[5] EURORDIS. The voice of rare disease patients in Europe. https://www.eurordis.org/about-rare-diseases. Accessed 4 Feb 2020.

[6] Langereis E.J. Wijburg F.A. Kuiper G., Meijer O.L.M. Failure to shorten the diagnostic delay in two ultra-orphan diseases (mucopolysaccharidosis types I and III): potential causes and implications. *Orphanet Journal of Rare Diseases*, 13:2, January 2018.

[7] Shire. The Global Challenge of Rare Disease Diagnosis. https://www.shire.com/-/media/shire/shireglobal/shirecom/pdffiles/patient/shire-diagnosis-initiative-pag-leaflet.pdf. Released in Feb 2016.

[8] EURORDIS Rare Barometer Graphic Report. Share and protect our health data! Rare disease patients' preferences on data sharing and protection. January 2020.

[9] Christoph Meinel and Harald Sack. *Digital communication: Communication, multimedia, security.* Springer Science & Business Media, 2014.

[10] Ettore Bolisani and Constantin Bratianu. *The Elusive Definition of Knowledge*, pages 1–22. 07 2018.

[11] Xander Wilcke, Peter Bloem, and Victor de Boer. The Knowledge Graph as the Default Data Model for Machine Learning. 2017.

[12] Mark Abraham Magumba, Peter Nabende, and Ernest Mwebaze. Ontology boosted deep learning for disease name extraction from Twitter messages. *Journal of Big Data*, 5(1):31, 2018.

[13] Andreas Blumauer Interview. Semantic Web Company: the Web as we know it has reached its limits due to a lack of semantics. `https://www.techcompanynews.com/semantic-web-company-web-know-reached-limits-due-lack-semantics/`. June 2018.

[14] Data Age 2025. The Digitization of the World From Edge to Core. `https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf`. November 2018.

[15] Jiawei Han, Micheline Kamber, and Jian Pei. *Data Mining: Concepts and Techniques, 3rd edition.* Morgan Kaufmann, 2011.

[16] Gartner Top 10 Data and Analytics Trends. `https://www.gartner.com/smarterwithgartner/gartner-top-10-data-analytics-trends/`. November 2019.

[17] Charu C. Aggarwal and ChengXiang Zhai, editors. *Mining Text Data.* Springer, 2012.

[18] Shu-Hsien Liao, Pei-Hui Chu, and Pei-Yuan Hsiao. Data mining techniques and applications - A decade review from 2000 to 2011. *Expert Syst. Appl.*, 39(12):11303–11311, 2012.

[19] Gary Miner, John Elder, Thomas Hill, Robert Nisbet, Dursun Delen, and Andrew Fast. *Practical Text Mining and Statistical Analysis for Non-Structured Text Data Applications.* Academic Press, Inc., USA, 1st edition, 2012.

[20] M. Maheswari. Text Mining: Survey on Techniques and Applications. *International Journal of Science and Research (IJSR)*, 6, 06 2017.

[21] Thomas K Landauer and Susan T Dumais. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, 104(2):211, 1997.

[22] Thomas K Landauer, Peter W Foltz, and Darrell Laham. An introduction to latent semantic analysis. *Discourse processes*, 25(2-3):259–284, 1998.

[23] Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391–407, 1990.

[24] Fernand Gobet. Vocabulary acquisition. 2015.

[25] Nathan Halko, Per-Gunnar Martinsson, and Joel A Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM review*, 53(2):217–288, 2011.

[26] Q&A with Tim Berners-Lee. *Businessweek*. Retrieved from `https://www.bloomberg.com/news/articles/2007-04-09/q-and-a-with-tim-berners-leebusinessweek-business-news-stock-market-and-financial-advice`. April 2007.

[27] Luís Miguel Oliveira Machado, Renato Rocha Souza, and Maria da Graça Simões. Semantic web or web of data? a diachronic study (1999 to 2017) of the publications of tim berners-lee and the world wide web consortium. *JASIST*, 70(7):701–714, 2019.

[28] Tim Berners-Lee, James Hendler, and Ora Lassila. The semantic web. *Scientific American*, 284(5):34–43, May 2001.

[29] Tim Berners-Lee. Semantic Web Road map. Website, 1998. http://www.w3.org/DesignIssues/Semantic.html.

[30] Eric Miller. An Introduction to the Resource Description Framework. *D-Lib Mag.*, 4(5), 1998.

[31] Brian McBride. The resource description framework (RDF) and its vocabulary description language RDFS. In Steffen Staab and Rudi Studer, editors, *Handbook on Ontologies*, International Handbooks on Information Systems, pages 51–66. Springer, 2004.

[32] D. Brickley and R. V. Guha. RDF Vocabulary Description Language 1.0: RDF Schema. 2004.

[33] D. L. McGuinness, F. Van Harmelen, et al. OWL Web Ontology Language Overview. *W3C recommendation*, 10(10), 2004.

[34] I. Horrocks, P. F. Patel-Schneider, H. Boley, S. Tabet, B. Grosof, M. Dean, et al. SWRL: A Semantic Web Rule Language combining OWL and RuleML. *W3C Member submission*, pages 21–79, 2004.

[35] E. Prud, A. Seaborne, et al. SPARQL Query Language for RDF. 2006.

[36] Frank Manola and Eric Miller. RDF Primer, 2004.

[37] S. Harris and A. Seaborne. SPARQL 1.1 Query Language. *W3C recommendation*, 2013.

[38] Jiexing Li, Arnd Christian König, Vivek R. Narasayya, and Surajit Chaudhuri. Robust estimation of resource consumption for SQL queries using statistical techniques. *PVLDB*, 5(11):1555–1566, 2012.

[39] Wentao Wu, Yun Chi, Shenghuo Zhu, Jun'ichi Tatemura, Hakan Hacigümüs, and Jeffrey F. Naughton. Predicting query execution time: Are optimizer cost models really unusable? In Christian S. Jensen, Christopher M. Jermaine, and Xiaofang Zhou, editors, *29th IEEE International Conference on Data Engineering, ICDE 2013, Brisbane, Australia, April 8-12, 2013*, pages 1081–1092. IEEE Computer Society, 2013.

[40] Qiuhua Tang, Zixiang Li, and Liping Zhang. An effective discrete artificial bee colony algorithm with idle time reduction techniques for two-sided assembly line balancing problem of type-ii. *Computers & Industrial Engineering*, 97:146–156, 2016.

[41] Davide Cerri and Alfonso Fuggetta. Open standards, open formats, and open source. *Journal of Systems and Software*, 80(11):1930–1937, 2007.

[42] The Open Definition. `http://opendefinition.org`, 2015.

[43] Christian Bizer, Tom Heath, and Tim Berners-Lee. Linked data - the story so far. *Int. J. Semantic Web Inf. Syst.*, 5(3):1–22, 2009.

[44] The Linked Open Data Cloud. `https://lod-cloud.net/`. Accessed 10 Feb 2020.

[45] Maria Pilar Escobar Esteban, Gustavo Candela Romero, Juan Trujillo, Manuel Marco Such, and Jesús Peral. Adding value to linked open data using a multidimensional model approach based on the RDF data cube vocabulary. *Computer Standards & Interfaces*, 68, 2020.

[46] G. Witmer. Dictionary of philosophy of mind-ontology. 11, 2004.

[47] John F. Sowa. *Knowledge representation: logical, philosophical, and computational foundations.* Brooks/Cole, 2000.

[48] Thomas R. Gruber. A translation approach to portable ontology specifications. page 27, 1993.

[49] Szymon Bobek, Grzegorz J. Nalepa, and Mateusz Slazynski. HEART-DROID - rule engine for mobile and context-aware expert systems. *Expert Systems*, 36(1), 2019.

[50] Ontotext Knowledge Hub. What are Ontologies? `https://www.ontotext.com/knowledgehub/fundamentals/what-are-ontologies/`. Accessed 10 Feb 2020.

[51] Aldo Gangemi and Valentina Presutti. Ontology design patterns. In *Handbook on ontologies*, pages 221–243. Springer, 2009.

[52] Heather Hedden. Controlled vocabularies, thesauri, and taxonomies. *The Indexer: The International Journal of Indexing*, 26(1):33–34, 2008.

[53] Luis Miguel Sintra Salvo Paiva. *Semantic relations extraction from unstructured information for domain ontologies enrichment.* PhD thesis, 2015.

[54] Celson Lima, Alain Zarli, and Graham Storer. Controlled vocabularies in the European construction sector: evolution, current developments, and future trends. In *Complex Systems Concurrent Engineering*, pages 565–574. Springer, 2007.

[55] WordNets in the World. *Global WordNet Association.* `http://globalwordnet.org/resources/wordnets-in-the-world/`. Retrieved 19 January 2020.

[56] George A. Miller. Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41, 1995.

[57] Emanuele Pianta, Luisa Bentivogli, and Christian Girardi. Multiwordnet: developing an aligned multilingual database. In *Proceedings of the First International Conference on Global WordNet*, January 2002.

[58] Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias,

editors, *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010, 17-23 May 2010, Valletta, Malta*. European Language Resources Association, 2010.

[59] N. Guarino. Formal Ontology in Information Systems. *Proceedings of the First International Conference*, 98:81–97, 1998.

[60] D. L. McGuinness. Ontologies come of age. *Mit Press*, 2005.

[61] Asunción Gómez-Pérez, Mariano Fernández-López, and Óscar Corcho. *Ontological Engineering: With Examples from the Areas of Knowledge Management, e-Commerce and the Semantic Web*. Advanced Information and Knowledge Processing. Springer, 2004.

[62] Zhu ChuangLu. Research on the semantic web reasoning technology. *AASRI Procedia*, 1:87–91, 2012.

[63] Franz Baader and Werner Nutt. Basic description logics. In Franz Baader, Diego Calvanese, Deborah L. McGuinness, Daniele Nardi, and Peter F. Patel-Schneider, editors, *The Description Logic Handbook: Theory, Implementation, and Applications*, pages 43–95. Cambridge University Press, 2003.

[64] Nikolaos Konstantinou and Dimitrios-Emmanuel Spanos. *Materializing the Web of Linked Data*. Springer, 2015.

[65] Rob Shearer, Boris Motik, and Ian Horrocks. Hermit: A highly-efficient OWL reasoner. In Catherine Dolbear, Alan Ruttenberg, and Ulrike Sattler, editors, *Proceedings of the Fifth OWLED Workshop on OWL: Experiences and Directions, collocated with the 7th International Semantic Web Conference (ISWC-2008), Karlsruhe, Germany, October 26-27, 2008*, volume 432 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2008.

[66] Volker Haarslev and Ralf Möller. Description of the RACER system and its applications. In Carole A. Goble, Deborah L. McGuinness, Ralf Möller, and Peter F. Patel-Schneider, editors, *Working Notes of the 2001 International Description Logics Workshop (DL-2001), Stanford, CA, USA, August 1-3, 2001*, volume 49 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2001.

[67] Evren Sirin, Bijan Parsia, Bernardo Cuenca Grau, Aditya Kalyanpur, and Yarden Katz. Pellet: A practical OWL-DL reasoner. *J. Web Semant.*, 5(2):51–53, 2007.

[68] Dmitry Tsarkov and Ian Horrocks. Fact++ description logic reasoner: System description. In Ulrich Furbach and Natarajan Shankar, editors, *Automated Reasoning, Third International Joint Conference, IJCAR 2006, Seattle, WA, USA, August 17-20, 2006, Proceedings*, volume 4130 of *Lecture Notes in Computer Science*, pages 292–297. Springer, 2006.

[69] Sunitha Abburu. A survey on ontology reasoners and comparison. *International Journal of Computer Applications*, 57(17), 2012.

[70] Deborah L McGuinness, Frank Van Harmelen, et al. Owl web ontology language overview. *W3C recommendation*, 10(10):2004, 2004.

[71] Robin Manhaeve, Sebastijan Dumancic, Angelika Kimmig, Thomas Demeester, and Luc De Raedt. Deepproblog: Neural probabilistic logic programming. In Katrien Beuls, Bart Bogaerts, Gianluca Bontempi, Pierre Geurts, Nick Harley, Bertrand Lebichot, Tom Lenaerts, Gilles Louppe, and Paul Van Eecke, editors, *Proceedings of the 31st Benelux Conference on Artificial Intelligence (BNAIC 2019) and the 28th Belgian Dutch Conference on Machine Learning (Benelearn 2019), Brussels, Belgium, November 6-8, 2019*, volume 2491 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2019.

[72] Hanene Agrebi, Afef Bahri, and Rafik Bouaziz. Fuzzy ontologies model for semantic web. 04 2010.

[73] Paulo Costa and Kathryn Laskey. Pr-owl: A framework for probabilistic ontologies. volume 150, pages 237–249, 05 2006.

[74] Fernando Bobillo and Umberto Straccia. Fuzzy ontology representation using OWL 2. *International Journal of Approximate Reasoning*, 2011.

[75] Mingxia Gao and Chunnian Liu. Extending OWL by fuzzy description logic. In *17th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'05)*, pages 6–pp. IEEE, 2005.

[76] Giorgos Stoilos and Giorgos B Stamou. Extending Fuzzy Description Logics for the Semantic Web. In *OWLED*, volume 258, 2007.

[77] Giorgos Stoilos, Giorgos Stamou, and Jeff Z Pan. Fuzzy extensions of OWL: Logical properties and reduction to fuzzy description logics. *International Journal of Approximate Reasoning*, 51(6):656–679, 2010.

[78] Valerie Cross and Shangye Chen. Fuzzy ontologies: State of the art revisited. In Guilherme A. Barreto and Ricardo Coelho, editors, *Fuzzy Information Processing*, pages 230–242, Cham, 2018. Springer International Publishing.

[79] Miroslav Vacura, Vojtěch Svátek, and Pavel Smrž. A pattern-based framework for uncertainty representation in ontologies. In Petr Sojka, Aleš Horák, Ivan Kopeček, and Karel Pala, editors, *Text, Speech and Dialogue*, pages 227–234, Berlin, Heidelberg, 2008. Springer Berlin Heidelberg.

[80] Mauro Mazzieri. A fuzzy RDF semantics to represent trust metadata. In *1st Workshop on Semantic Web Applications and Perspectives (SWAP2004)*, volume 1, pages 83–89, 2004.

[81] Lotfi Asker Zadeh. The concept of a linguistic variable and its application to approximate reasoning. In *Learning systems and intelligent robots*, pages 1–10. Springer, 1974.

[82] Rachel Pereira, Ivan Ricarte, and Fernando Gomide. Fuzzy relational ontological model in information search systems. *Fuzzy logic and the semantic Web, capturing intelligence*, pages 395–412, 2006.

[83] Paulo Gottgtroy, Nikola Kasabov, Stephen MacDonell, and E Sanchez. Evolving ontologies for intelligent decision support. *Fuzzy logic and the semantic Web, capturing intelligence*, pages 415–440, 2006.

[84] Giorgos Stoilos, Nikos Simou, Giorgos Stamou, and Stefanos Kollias. Uncertainty and the semantic web. *Intelligent Systems, IEEE*, 21:84 – 87, 10 2006.

[85] Umberto Straccia. A fuzzy description logic for the semantic web. In *Capturing Intelligence*, volume 1, pages 73–90. Elsevier, 2006.

[86] Fernando Bobillo and Umberto Straccia. Towards a crisp representation of fuzzy description logics under łukasiewicz semantics. In *International Symposium on Methodologies for Intelligent Systems*, pages 309–318. Springer, 2008.

[87] Fernando Bobillo, Miguel Delgado, and Juan Gómez-Romero. Optimizing the Crisp Representation of the Fuzzy Description Logic SROIQ. In *Uncertainty Reasoning for the Semantic Web I*, pages 189–206. Springer, 2006.

[88] Fuzzy OWL 2. Protégé plug-in. `http://www.umbertostraccia.it/cs/software/FuzzyOWL/index.html#plug-in`. Accessed 12 Feb 2020.

[89] Fernando Bobillo and Umberto Straccia. fuzzyDL: An expressive fuzzy description logic reasoner. In *2008 IEEE International Conference on Fuzzy Systems (IEEE World Congress on Computational Intelligence)*, pages 923–930. IEEE, 2008.

[90] Fernando Bobillo, Miguel Delgado, and Juan Gómez-Romero. DeLorean: A reasoner for fuzzy OWL 2. *Expert Systems with Applications*, 39(1):258–272, 2012.

[91] Rommel N. Carvalho, Kathryn B. Laskey, and Paulo C.G. Costa. Pr-owl – a language for defining probabilistic ontologies. *International Journal of Approximate Reasoning*, 91:56 – 79, 2017.

[92] Rommel Carvalho. *Probabilistic Ontology: Representation and Modeling Methodology*. PhD thesis, 01 2011.

[93] UnBBayes. Java tool supporting PR-OWL 2. `https://sourceforge.net/projects/unbbayes/`. Accessed 12 Feb 2020.

[94] GO Blog. Introducing the knowledge graph: thing, not strings. *Introducing the Knowledge Graph: things, not strings*, 2012.

[95] Dieter Fensel, Umutcan Şimşek, Kevin Angele, Elwin Huaman, Elias Kärle, Oleksandra Panasiuk, Ioan Toma, Jürgen Umbrich, and Alexander Wahler. *Knowledge Graphs: Methodology, Tools and Selected Use Cases*. Springer Nature, 2020.

[96] Medium Data Science. Where Ontologies End and Knowledge Graphs Begin. `https://medium.com/predict/where-ontologies-end-and-knowledge-graphs-begin-6fe0cdede1ed`. Accessed 14 Feb 2020.

[97] Ontotext. What is a Knowledge Graph? `https://www.ontotext.com/knowledgehub/fundamentals/what-is-a-knowledge-graph/`. Accessed 14 Feb 2020.

[98] Connecting the Dots: Using AI & Knowledge Graphs to Identify Investment Opportunities. `https://towardsdatascience.com/knowledge-graphs-in-investing-733ab34abe`. Accessed 27 Feb 2020.

[99] Heiko Paulheim. Knowledge graph refinement: A survey of approaches and evaluation methods. *Semantic web*, 8(3):489–508, 2017.

[100] M Kroetsch and G Weikum. Special issue on knowledge graphs. *Journal of Web Semantics*, 37(38):53–54, 2016.

[101] Michael Färber, Frederic Bartscherer, Carsten Menne, and Achim Rettinger. Linked data quality of dbpedia, freebase, opencyc, wikidata, and yago. *Semantic Web*, 9(1):77–129, 2018.

[102] Jay Pujara, Hui Miao, Lise Getoor, and William Cohen. Knowledge graph identification. In *International Semantic Web Conference*, pages 542–557. Springer, 2013.

[103] Lisa Ehrlinger and Wolfram Wöß. Towards a Definition of Knowledge Graphs. *SEMANTiCS (Posters, Demos, SuCCESS)*, 48, 2016.

[104] Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 1247–1250, 2008.

[105] Christian Bizer, Jens Lehmann, Georgi Kobilarov, Sören Auer, Christian Becker, Richard Cyganiak, and Sebastian Hellmann. DBpedia - A crystallization point for the Web of Data. *Journal of web semantics*, 7(3):154–165, 2009.

[106] Johannes Hoffart, Fabian M Suchanek, Klaus Berberich, Edwin Lewis-Kelham, Gerard De Melo, and Gerhard Weikum. YAGO2: exploring and querying world knowledge in time, space, context, and many languages. In *Proceedings of the 20th international conference companion on World wide web*, pages 229–232, 2011.

[107] George A Miller. *WordNet: An electronic lexical database.* MIT press, 1998.

[108] Gartner Hype Cycle for Emerging Technologies 2019. `https://www.gartner.com/smarterwithgartner/5-trends-appear-on-the-gartner-hype-cycle-for-emerging-technologies-2019/`. Accessed 12 Feb 2020.

[109] Lynn Marie Schriml, Cesar Arze, Suvarna Nadendla, Yu-Wei Wayne Chang, Mark Mazaitis, Victor Felix, Gang Feng, and Warren Alden Kibbe. Disease Ontology: a backbone for disease semantic integration. *Nucleic acids research*, 40(D1):D940–D946, 2012.

[110] Warren A. Kibbe, Cesar Arze, Victor Felix, Elvira Mitraka, Evan Bolton, Gang Fu, Christopher J. Mungall, Janos X. Binder, James Malone, Drashtti Vasant, Helen Parkinson, and Lynn M. Schriml. Disease Ontology 2015 update: an expanded and updated database of human diseases

for linking biomedical knowledge through disease data. *Nucleic Acids Research*, 43(D1):D1071–D1078, 10 2014.

[111] Susan M Bello, Mary Shimoyama, Elvira Mitraka, Stanley JF Laulederkind, Cynthia L Smith, Janan T Eppig, and Lynn M Schriml. Disease Ontology: improving and unifying disease annotations across species. *Disease models & mechanisms*, 11(3):dmm032839, 2018.

[112] Trevor Cohen and Dominic Widdows. *Geometric Representations in Biomedical Informatics: Applications in Automated Text Analysis*, pages 99–139. 12 2014.

[113] Ana Rath, Annie Olry, Ferdinand Dhombres, Maja Miličić Brandt, Bruno Urbero, and Segolene Ayme. Representation of rare diseases in health information systems: the Orphanet approach to serve a wide range of end users. *Human mutation*, 33(5):803–808, 2012.

[114] Drashtti Vasant, Laetitia Chanas, James Malone, Marc Hanauer, Annie Olry, Simon Jupp, Peter N Robinson, Helen Parkinson, and Ana Rath. Ordo: An ontology connecting rare disease, epidemiology and genetic data. In *Proceedings of ISMB*, volume 30, 2014.

[115] RD-Connect. Ontologies in Rare Disease Registries. `https://rd-connect.eu/what-we-do/data-linkage/ontologies-in-rare-disease-registries/`. Accessed 12 Feb 2020.

[116] Sebastian Köhler, Nicole A. Vasilevsky, Mark Engelstad, Erin Foster, Julie McMurry, Ségolène Aymé, Gareth Baynam, Susan M. Bello, Cornelius F. Boerkoel, Kym M. Boycott, Michael Brudno, Orion J. Buske, Patrick F. Chinnery, Valentina Cipriani, Laureen E. Connell, Hugh J.S. Dawkins, Laura E. DeMare, Andrew D. Devereau, Bert B.A. de Vries, Helen V. Firth, Kathleen Freson, Daniel Greene, Ada Hamosh, Ingo Helbig, Courtney Hum, Johanna A. Jähn, Roger James, Roland Krause, Stanley J. F. Laulederkind, Hanns Lochmüller, Gholson J. Lyon, Soichi Ogishima, Annie Olry, Willem H. Ouwehand, Nikolas Pontikos, Ana Rath, Franz Schaefer, Richard H. Scott, Michael Segal, Panagiotis I. Sergouniotis, Richard Sever, Cynthia L. Smith, Volker Straub, Rachel Thompson, Catherine Turner, Ernest Turro, Marijcke W.M. Veltman, Tom Vulliamy, Jing Yu, Julie von Ziegenweidt, Andreas Zankl, Stephan Züchner, Tomasz Zemojtel, Julius O.B. Jacobsen, Tudor Groza, Damian Smedley, Christopher J. Mungall, Melissa Haendel, and Peter N. Robinson. The Human Phenotype Ontology in 2017. *Nucleic Acids Research*, 45(D1):D865–D876, 11 2016.

[117] What is HOOM (The ORDO-HOOM Ontological Module)? `http://www.orphadata.org/cgi-bin/img/PDF/WhatIsHOOM.pdf`. September 2019.

[118] Ada Hamosh, Alan F Scott, Joanna Amberger, Carol Bocchini, David Valle, and Victor A McKusick. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic acids research*, 30(1):52–55, 2002.

[119] Tudor Groza, Sebastian Köhler, Dawid Moldenhauer, Nicole Vasilevsky, Gareth Baynam, Tomasz Zemojtel, Lynn Marie Schriml, Warren Alden Kibbe, Paul N. Schofield, Tim Beck, Drashtti Vasant, Anthony J. Brookes, Andreas Zankl, Nicole L. Washington, Christopher J. Mungall, Suzanna E. Lewis, Melissa A. Haendel, Helen Parkinson, and Peter N. Robinson. The human phenotype ontology: Semantic unification of common and rare disease. *The American Journal of Human Genetics*, 97(1):111 – 124, 2015.

[120] Sebastian Köhler, Marcel H. Schulz, Peter Krawitz, Sebastian Bauer, Sandra Dölken, Claus E. Ott, Christine Mundlos, Denise Horn, Stefan Mundlos, and Peter N. Robinson. Clinical diagnostics in human genetics with semantic similarity searches in ontologies. *The American Journal of Human Genetics*, 85(4):457 – 464, 2009.

[121] Jr. Kahn, Charles E. Integrating ontologies of rare diseases and radiological diagnosis. *Journal of the American Medical Informatics Association*, 22(6):1164–1168, 04 2015.

[122] Xhemal Zenuni, Bujar Raufi, Florie Ismaili, and Jaumin Ajdari. State of the Art of Semantic Web for Healthcare. *Procedia - Social and Behavioral Sciences*, 195:1990–1998, 07 2015.

[123] Charles Elkan. *Predictive analytics and data mining.* University of California, 2013.

[124] Zhiheng Huang, Wei Xu, and Kai Yu. Bidirectional LSTM-CRF models for sequence tagging. *arXiv preprint arXiv:1508.01991*, 2015.

[125] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642, 2013.

[126] Yoon Kim. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*, 2014.

[127] Danqi Chen and Christopher D Manning. A fast and accurate dependency parser using neural networks. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 740–750, 2014.

[128] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[129] Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*, 73:1–15, 2018.

[130] Anh Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 427–436, 2015.

[131] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Universal adversarial perturbations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1765–1773, 2017.

[132] Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z Berkay Celik, and Ananthram Swami. The limitations of deep learning in adversarial settings. In *2016 IEEE European symposium on security and privacy (EuroS&P)*, pages 372–387. IEEE, 2016.

[133] Robin Jia and Percy Liang. Adversarial examples for evaluating reading comprehension systems. *arXiv preprint arXiv:1707.07328*, 2017.

[134] David Gunning. Explainable Artificial Intelligence (XAI). *Defense Advanced Research Projects Agency (DARPA), nd Web*, 2, 2017.

[135] Leilani H Gilpin, David Bau, Ben Z Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. Explaining explanations: An Overview of Interpretability of Machine Learning. In *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*, pages 80–89. IEEE, 2018.

[136] Sherin Mary. *Explainable Artificial Intelligence Applications in NLP, Biomedical, and Malware Classification: A Literature Review*, pages 1269–1292. 07 2019.

[137] Rishi Chhatwal, Peter Gronvall, Nathaniel Huber-Fliflet, Robert Keeling, Jianping Zhang, and Haozhen Zhao. Explainable Text Classification in Legal Document Review A Case Study of Explainable Predictive Coding. In *2018 IEEE International Conference on Big Data (Big Data)*, pages 1905–1911. IEEE, 2018.

[138] Jin Hu. Explainable Deep Learning for Natural Language Processing, 2018.

[139] IBM Watson Explorer. `https://www.ibm.com/it-it/products/watson-explorer`. Accessed 21 Feb 2020.

[140] Robert E Hoyt, Dallas H Snider, Carla J Thompson, and Sarita Mantravadi. IBM Watson analytics: automating visualization, descriptive, and predictive statistics. *JMIR public health and surveillance*, 2(2):e157, 2016.

[141] Helena Ahonen, Oskari Heinonen, Mika Klemettinen, and A Inkeri Verkamo. Applying Data Mining Techniques for Descriptive Phrase Extraction in Digital Document Collections. In *Proceedings IEEE International Forum on Research and Technology Advances in Digital Libraries-ADL'98-*, pages 2–11. IEEE, 1998.

[142] Bing Liu and Lei Zhang. A survey of opinion mining and sentiment analysis. In *Mining text data*, pages 415–463. Springer, 2012.

[143] Minqing Hu and Bing Liu. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177, 2004.

[144] Ana-Maria Popescu and Orena Etzioni. Extracting product features and opinions from reviews. In *Natural language processing and text mining*, pages 9–28. Springer, 2007.

[145] Li Zhuang, Feng Jing, and Xiao-Yan Zhu. Movie review mining and summarization. In *Proceedings of the 15th ACM international conference on Information and knowledge management*, pages 43–50, 2006.

[146] Ivan Titov and Ryan McDonald. Modeling online reviews with multi-grain topic models. In *Proceedings of the 17th international conference on World Wide Web*, pages 111–120, 2008.

[147] Chenghua Lin and Yulan He. Joint sentiment/topic model for sentiment analysis. In *Proceedings of the 18th ACM conference on Information and knowledge management*, pages 375–384, 2009.

[148] Qiaozhu Mei, Xu Ling, Matthew Wondra, Hang Su, and ChengXiang Zhai. Topic sentiment mixture: modeling facets and opinions in weblogs. In *Proceedings of the 16th international conference on World Wide Web*, pages 171–180, 2007.

[149] Samuel Brody and Noemie Elhadad. An unsupervised aspect-sentiment model for online reviews. In *Human language technologies: The 2010 annual conference of the North American chapter of the association for computational linguistics*, pages 804–812. Association for Computational Linguistics, 2010.

[150] Wayne Xin Zhao, Jing Jiang, Hongfei Yan, and Xiaoming Li. Jointly modeling aspects and opinions with a MaxEnt-LDA hybrid. In *Proceedings of the 2010 conference on empirical methods in natural language processing*, pages 56–65. Association for Computational Linguistics, 2010.

[151] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[152] Ivan Titov and Ryan McDonald. A joint model of text and aspect ratings for sentiment summarization. In *proceedings of ACL-08: HLT*, pages 308–316, 2008.

[153] Yue Lu, ChengXiang Zhai, and Neel Sundaresan. Rated aspect summarization of short comments. In *Proceedings of the 18th international conference on World wide web*, pages 131–140, 2009.

[154] Hongning Wang, Yue Lu, and Chengxiang Zhai. Latent aspect rating analysis on review text data: a rating regression approach. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 783–792, 2010.

[155] Yohan Jo and Alice H Oh. Aspect and sentiment unification model for online review analysis. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 815–824, 2011.

[156] Sherin Mary. *Explainable Artificial Intelligence Applications in NLP, Biomedical, and Malware Classification: A Literature Review*, pages 1269–1292. 07 2019.

[157] Wojciech Samek, Thomas Wiegand, and Klaus-Robert Müller. Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *arXiv preprint arXiv:1708.08296*, 2017.

[158] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

[159] Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38, 2019.

[160] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

[161] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3145–3153. JMLR. org, 2017.

[162] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in neural information processing systems*, pages 4765–4774, 2017.

[163] Jörg Landthaler, Ingo Glaser, and Florian Matthes. Towards Explainable Semantic Text Matching. In *JURIX*, pages 200–204, 2018.

[164] Bernhard Waltl, Georg Bonczek, Elena Scepankova, and Florian Matthes. Semantic types of legal norms in German laws: classification and analysis using local linear explanations. *Artificial Intelligence and Law*, 27(1):43–71, 2019.

[165] Omar Zaidan, Jason Eisner, and Christine Piatko. Using "annotator rationales" to improve machine learning for text categorization. In *Human language technologies 2007: The conference of the North American chapter of the association for computational linguistics; proceedings of the main conference*, pages 260–267, 2007.

[166] David Martens and Foster Provost. Explaining data-driven document classifications. *Mis Quarterly*, 38(1):73–100, 2014.

[167] Ye Zhang, Iain Marshall, and Byron C Wallace. Rationale-augmented convolutional neural networks for text classification. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing.*

Conference on Empirical Methods in Natural Language Processing, volume 2016, page 795. NIH Public Access, 2016.

[168] Tao Lei, Regina Barzilay, and Tommi Jaakkola. Rationalizing neural predictions. *arXiv preprint arXiv:1606.04155*, 2016.

[169] Hui Liu, Qingyu Yin, and William Yang Wang. Towards explainable nlp: A generative explanation framework for text classification. *arXiv preprint arXiv:1811.00196*, 2018.

[170] Eric Wallace, Jens Tuyls, Junlin Wang, Sanjay Subramanian, Matt Gardner, and Sameer Singh. AllenNLP Interpret: A Framework for Explaining Predictions of NLP Models. *arXiv preprint arXiv:1909.09251*, 2019.

[171] Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. AllenNLP: A deep semantic natural language processing platform. *arXiv preprint arXiv:1803.07640*, 2018.

[172] Javid Ebrahimi, Anyi Rao, Daniel Lowd, and Dejing Dou. Hotflip: White-box adversarial examples for text classification. *arXiv preprint arXiv:1712.06751*, 2017.

[173] Shi Feng, Eric Wallace, II Grissom, Mohit Iyyer, Pedro Rodriguez, Jordan Boyd-Graber, et al. Pathologies of neural models make interpretations difficult. *arXiv preprint arXiv:1804.07781*, 2018.

[174] Ada Brunstein. Annotation guidelines for answer types. *LDC2005T33, Linguistic Data Consortium, Philadelphia*, 2002.

[175] Alan Ritter, Sam Clark, Oren Etzioni, et al. Named entity recognition in tweets: an experimental study. In *Proceedings of the conference on empirical methods in natural language processing*, pages 1524–1534. Association for Computational Linguistics, 2011.

[176] Mohamed Amir Yosef, Sandro Bauer, Johannes Hoffart, Marc Spaniol, and Gerhard Weikum. Hyena: Hierarchical type classification for entity names. In *Proceedings of COLING 2012: Posters*, pages 1361–1370, 2012.

[177] Named Entity Recognition for Unstructured Documents. `https://medium.com/@dudsdu/named-entity-recognition-for-unstructured-documents-c325d47c7e3a`. Accessed 4 Mar 2020.

[178] Nal Kalchbrenner, Lasse Espeholt, Karen Simonyan, Aaron van den Oord, Alex Graves, and Koray Kavukcuoglu. Neural machine translation in linear time. *arXiv preprint arXiv:1610.10099*, 2016.

[179] Jason Lee, Kyunghyun Cho, and Thomas Hofmann. Fully character-level neural machine translation without explicit segmentation. *Transactions of the Association for Computational Linguistics*, 5:365–378, 2017.

[180] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017.

[181] Giacomo Domeniconi, Gianluca Moro, Roberto Pasolini, and Claudio Sartori. A Study on Term Weighting for Text Categorization: A Novel Supervised Variant of tf. idf. In *DATA*, pages 26–37, 2015.

[182] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

[183] Vikas Raunak. Effective dimensionality reduction for word embeddings. *CoRR*, abs/1708.03629, 2017.

[184] Priscilla E Greenwood and Michael S Nikulin. *A guide to chi-squared testing*, volume 280. John Wiley & Sons, 1996.

[185] Lior Rokach and Oded Maimon. Clustering methods. In *Data mining and knowledge discovery handbook*, pages 321–352. Springer, 2005.

[186] Joe H Ward Jr. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301):236–244, 1963.

[187] Ryota Suzuki and Hidetoshi Shimodaira. Pvclust: an R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics*, 22(12):1540–1542, 04 2006.

[188] Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and Philip S Yu. A comprehensive survey on graph neural networks. *arXiv preprint arXiv:1901.00596*, 2019.

[189] Alexander Maedche and Steffen Staab. Ontology learning for the semantic web. *IEEE Intelligent systems*, 16(2):72–79, 2001.

[190] Muhammad Nabeel Asim, Muhammad Wasim, Muhammad Usman Ghani Khan, Waqar Mahmood, and Hafiza Mahnoor Abbasi. A survey of ontology learning techniques and applications. *Database*, 2018, 2018.

[191] P Buitelaar, P Cimiano, and B Magnini. Ontology learning from text: An overview. ontology learning from text: Methods, evaluation and applications. *Frontiers in Artificial Intelligence and Applications Series*, 123, 2005.

[192] Xing Jiang and Ah-Hwee Tan. CRCTOL: A semantic-based domain ontology learning system. *Journal of the American Society for Information Science and Technology*, 61(1):150–168, 2010.

[193] Andrew Hippisley, David Cheng, and Khurshid Ahmad. The head-modifier principle and multilingual term extraction. *Natural Language Engineering*, 11(2):129–157, 2005.

[194] Alexandre Agustini, Pablo Gamallo, and Gabriel P Lopes. Selection restrictions acquisition for parsing improvement. In *International Conference on Applications of Prolog*, pages 129–143. Springer, 2001.

[195] Pablo Gamallo, Alexandre Agustini, and Gabriel P Lopes. Learning subcategorisation information to model a grammar with" co-restrictions". 2003.

[196] MAEF Belal, H Abdel-Galil, and YM Saber. Ontology extraction from text: Related works between arabic and english languages. *Int. J*, 4(8), 2016.

[197] David Sanchez and Antonio Moreno. Creating ontologies from web documents. *Recent advances in artificial intelligence research and development. IOS Press*, 113:11–18, 2004.

[198] Alvaro L Fraga and Marcela Vegetti. Semi-automated ontology generation process from industrial product data standards. In *III Simposio Argentino de Ontologías y sus Aplicaciones (SAOA)-JAIIO 46 (Córdoba, 2017).*, 2017.

[199] Neha Kaushik and Niladri Chatterjee. Automatic relationship extraction from agricultural text for ontology construction. *Information processing in agriculture*, 5(1):60–73, 2018.

[200] Katerina Frantzi, Sophia Ananiadou, and Hideki Mima. Automatic recognition of multi-word terms:. the c-value/nc-value method. *International journal on digital libraries*, 3(2):115–130, 2000.

[201] E Milios, Y Zhang, B He, and L Dong. Automatic term extraction and document similarity in special text corpora. In *Proceedings of the sixth*

conference of the pacific association for computational linguistics, pages 275–284. Citeseer, 2003.

[202] Roberto Navigli, Paola Velardi, and Aldo Gangemi. Ontology learning and its application to automated terminology translation. *IEEE Intelligent systems*, 18(1):22–31, 2003.

[203] Bouchra Frikh, Ahmed Said Djaanfar, and Brahim Ouhbi. A Hybrid Method for Domain Ontology Construction from the Web. In *KEOD*, pages 285–292, 2011.

[204] S Suresu and M Elamparithi. Probabilistic relational concept extraction in ontology learning. *Int. J. Inform. Technol*, 2, 2016.

[205] Monika Rani, Amit Kumar Dhar, and OP Vyas. Semi-automatic terminology ontology learning based on topic modeling. *Engineering Applications of Artificial Intelligence*, 63:108–125, 2017.

[206] Lobna Karoui, Marie-Aude Aufaure, and Nacera Bennacer. Contextual Concept Discovery Algorithm. In *FLAIRS Conference*, pages 460–465, 2007.

[207] Ishara Sandun, Sagara Sumathipala, and Gamage Upeksha Ganegoda. Self-evolving disease ontology for medical domain based on web. *International Journal of Fuzzy Logic and Intelligent Systems*, 17(4):307–314, 2017.

[208] Hermine Njike Fotzo and Patrick Gallinari. Learning generalization/-specialization relations between concepts-application for automatically building thematic document hierarchies. In *RIAO*, pages 143–155, 2004.

[209] Euthymios Drymonas, Kalliopi Zervanou, and Euripides GM Petrakis. Unsupervised ontology acquisition from plain texts: the OntoGain system. In *International Conference on Application of Natural Language to Information Systems*, pages 277–287. Springer, 2010.

[210] David Faure and Claire Nédellec. A corpus-based conceptual clustering method for verb frames and ontology acquisition. In *LREC workshop on adapting lexical and corpus resources to sublanguages and applications*, volume 707, page 30. Citeseer, 1998.

[211] Alexander Maedche and Steffen Staab. Measuring similarity between ontologies. In *International Conference on Knowledge Engineering and Knowledge Management*, pages 251–263. Springer, 2002.

[212] Philipp Cimiano and Steffen Staab. Learning concept hierarchies from text with a guided agglomerative clustering algorithm. In *Proceedings of the ICML 2005 Workshop on Learning and Extending Lexical Ontologies with Machine Learning Methods*, 2005.

[213] Luis Miguel Sintra Salvo Paiva. *Semantic relations extraction from unstructured information for domain ontologies enrichment.* PhD thesis, 2015.

[214] Rihab Idoudi, Karim Saheb Ettabaa, Basel Solaiman, and Najla Mnif. Association rules based ontology enrichment. *International journal of web applications*, 8(1):16–25, 2016.

[215] David S Batista, Bruno Martins, and Mário J Silva. Semi-supervised bootstrapping of relationship extractors with distributional semantics. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 499–504, 2015.

[216] Nanda Kambhatla. Combining lexical, syntactic, and semantic features with maximum entropy models for extracting relations. In *Proceedings of the ACL 2004 on Interactive poster and demonstration sessions*, pages 22–es, 2004.

[217] Enrique Alfonseca, Katja Filippova, Jean-Yves Delort, and Guillermo Garrido. Pattern Learning for Relation Extraction with Hierarchical Topic Models. 2012.

[218] Rinaldo Lima, Bernard Espinasse, Hilário Oliveira, Rafael Ferreira, Luciano Cabral, Fred Freitas, Renê Gadelha, et al. An inductive logic programming-based approach for ontology population from the web. In *International Conference on Database and Expert Systems Applications*, pages 319–326. Springer, 2013.

[219] Nathalie Aussenac-Gilles, Sylvie Despres, and Sylvie Szulman. The terminae method and platform for ontology engineering from texts. volume 167, pages 199–223, 06 2008.

[220] Philipp Cimiano and Johanna Völker. Text2Onto. In *International conference on application of natural language to information systems*, pages 227–238. Springer, 2005.

[221] Blaz Fortuna, Marko Grobelnik, and Dunja Mladenic. OntoGen: semi-automatic ontology editor. In *Symposium on Human Interface and the Management of Information*, pages 309–318. Springer, 2007.

[222] Xing Jiang and Ah-Hwee Tan. CRCTOL: A semantic-based domain ontology learning system. *Journal of the American Society for Information Science and Technology*, 61(1):150–168, 2010.

[223] Pablo N Mendes, Max Jakob, Andrés García-Silva, and Christian Bizer. DBpedia spotlight: shedding light on the web of documents. In *Proceedings of the 7th international conference on semantic systems*, pages 1–8, 2011.

[224] Theerayut Thongkrau and Pattarachai Lalitrojwong. Ontopop: An ontology population system for the semantic web. *IEICE TRANSACTIONS on Information and Systems*, 95(4):921–931, 2012.

[225] Aldo Gangemi, Valentina Presutti, Diego Reforgiato Recupero, Andrea Giovanni Nuzzolese, Francesco Draicchio, and Misael MongiovÃ¬. Semantic Web Machine Reading with FRED. *Semantic Web*, 8(6):873–893, 2017.

[226] Andrés Montoyo, Rafael Muñoz, and Elisabeth Métais. *Natural Language Processing and Information Systems: 10th International Conference on Applications of Natural Language to Information Systems, NLDB 2005, Alicante, Spain, June 15-17, Proceedings*, volume 3513. Springer Science & Business Media, 2005.

[227] Aldo Gangemi. A comparison of knowledge extraction tools for the semantic web. In *Extended semantic web conference*, pages 351–366. Springer, 2013.

[228] Orphanet. Idiopathic achalasia. `https://www.orpha.net/consor/cgi-bin/OC_Exp.php?lng=EN&Expert=930`. Accessed 4 Mar 2020.

[229] National Organization for Rare Disorder. Achalasia. `https://rarediseases.org/rare-diseases/achalasia/`. Accessed 4 Mar 2020.

[230] Dhyanesh A Patel, Brian M Lappas, and Michael F Vaezi. An overview of achalasia and its subtypes. *Gastroenterology & hepatology*, 13(7):411, 2017.

[231] Wikipedia. Achalasia. `https://en.wikipedia.org/wiki/Esophageal_achalasia`. Accessed 4 Mar 2020.

[232] Ifeanyi I Momodu and Jason M Wallen. Achalasia. 2019.

[233] Achalasia Patient Organizations, Orphanet. `https://bit.ly/2vUYN70`. Accessed 9 Mar 2020.

[234] DBPedia and Wikidata. https://meta.wikimedia.org/wiki/Wikidata/Notes/DBpedia_and_Wikidata. Accessed 10 Mar 2020.

[235] Brandon Pincombe. Comparison of human and latent semantic analysis (LSA) judgements of pairwise document similarities for a news corpus. Technical report, DEFENCE SCIENCE AND TECHNOLOGY ORGANISATION SALISBURY (AUSTRALIA), 2004.

[236] Michael D Lee, Brandon Pincombe, and Matthew Welsh. An empirical evaluation of models of text document similarity. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 27, 2005.