UNIVERSIDAD POLITÉCNICA DE MADRID

ESCUELA TÉCNICA SUPERIOR DE INGENIEROS DE
TELECOMUNICACIÓN

SENTIMENT AND EMOTION ANALYSIS IN SOCIAL
NETWORKS: MODELING AND LINKING DATA, AFFECTS
AND PEOPLE

TESIS DOCTORAL

J. FERNANDO SÁNCHEZ RADA
Ingeniero de Telecomunicación

2019

UNIVERSIDAD POLITÉCNICA DE MADRID

ESCUELA TÉCNICA SUPERIOR DE INGENIEROS DE
TELECOMUNICACIÓN

SENTIMENT AND EMOTION ANALYSIS IN SOCIAL
NETWORKS: MODELING AND LINKING DATA, AFFECTS
AND PEOPLE

TESIS DOCTORAL

J. FERNANDO SÁNCHEZ RADA
Ingeniero de Telecomunicación

2019

DEPARTAMENTO DE INGENIERÍA DE SISTEMAS
TELEMÁTICOS

ESCUELA TÉCNICA SUPERIOR DE INGENIEROS DE
TELECOMUNICACIÓN

UNIVERSIDAD POLITÉCNICA DE MADRID

SENTIMENT AND EMOTION ANALYSIS IN SOCIAL
NETWORKS: MODELING AND LINKING DATA, AFFECTS
AND PEOPLE

AUTOR:

J. FERNANDO SÁNCHEZ RADA
Ingeniero de Telecomunicación

TUTOR:

CARLOS A. IGLESIAS
Doctor Ingeniero de Telecomunicación

2019

Tribunal nombrado por el Magfco. y Excmo. Sr. Rector de la Universidad Politécnica de Madrid, el día _____ de _____ de _____.

Presidente: _____

Vocal: _____

Vocal: _____

Vocal: _____

Secretario: _____

Suplente: _____

Suplente: _____

Realizado el acto de defensa y lectura de la Tesis el día _____ de _____ de _____. en la E.T.S.I. Telecomunicación habiendo obtenido la calificación de _____.

EL PRESIDENTE

LOS VOCALES

EL SECRETARIO

A mi familia y amigos.

Resumen

El objetivo principal de esta tesis doctoral es mejorar el análisis de sentimientos y emociones de texto en redes sociales, aunando técnicas de procesamiento de lenguaje natural, datos enlazados y análisis de redes sociales. La investigación se divide en tres partes muy diferenciadas.

Primero, se desarrolló un vocabulario semántico para describir emociones y procesos de análisis de sentimientos, alineado con la ontología de “procedencia” PROV-O. Este vocabulario permite seguir un enfoque de datos enlazados en el análisis de emociones, tanto en la anotación de recursos (datasets y lexicons), como en la publicación de servicios semánticos de análisis de emociones. Asimismo, se extendió el vocabulario de referencia para opiniones, Marl, para alinearlo con Prov-O.

En segundo lugar, se han modelado los diferentes componentes de los servicios de análisis de sentimientos y emociones, así como los requisitos para crear servicios abiertos, interoperables y que se puedan combinar para lograr análisis avanzados. El resultado es un marco de desarrollo y modelado de servicios, enfocado en la modularidad. Además, se ha desarrollado una implementación de referencia que permite a crear y publicar servicios de análisis de sentimientos y emociones.

En tercer lugar, se ha caracterizado el contexto social, que es el conjunto de información en una red social que complementa al mensaje, y que puede ser utilizado para mejorar el análisis de sentimientos del mensaje. También se ha desarrollado una taxonomía de enfoques de análisis de sentimientos basada en la forma en que el contexto social es construido y utilizado en el análisis. Seguidamente, se han investigado modelos de análisis de sentimientos que utilizan contexto social enriquecido mediante análisis de redes sociales. Por último, para explorar el potencial de las diferentes teorías sociales para el análisis de sentimientos se ha desarrollado una plataforma de simulación social, en la que se han implementado varios modelos de propagación de rumores y emociones.

Abstract

The main goal of this thesis is to improve sentiment and emotion analysis of text in social media through a combination of natural language processing, linked data and social network analysis. To achieve this goal, we have divided our research into three parts.

First, we developed a semantic vocabulary to describe emotions, emotion models and emotion analysis activities. This vocabulary enables a linked data approach to emotion analysis, including in the annotation and processing of resources (e.g., datasets and lexicons), and the development of public semantic emotion analysis services. We also extended the most popular vocabulary for opinions and sentiment, Marl, to include concepts of sentiment analysis activities.

Secondly, we modeled the different components in a sentiment or emotion analysis service, as well as the requirements to create public and interoperable services that can be composed to produce advanced analyses. The result is a framework to model and develop modular services. We also developed a reference implementation of this framework, which can be used by researchers and developers to create and publish new sentiment and emotion analysis services.

Thirdly, we studied and formalized the concept of social context, which is the information in a social network that accompanies a text message and can be used to improve the analysis of said text. We also developed a taxonomy of approaches to sentiment analysis based on how they gather social context and how they exploit it in the analysis. In addition to characterizing social context, we investigated several models of sentiment analysis that enrich social context through social network analysis. Lastly, we developed a social simulation platform, in which we modelled several rumor and emotion propagation behaviors.

Contents

Resumen	I
Abstract	III
Contents	V
Reader's guide	1
1 Introduction	3
1.1 Motivation	4
1.2 Background	8
1.2.1 Sentiment and Emotion Analysis	8
1.2.2 Linked Data for Natural Language Processing (NLP)	10
1.2.3 Emotion Representation	16
1.2.4 Contextual information for sentiment analysis	19
1.2.5 Social Network Analysis and Community Detection	20
1.3 Hypotheses	21
1.4 Objectives	21
1.5 Document outline	22
2 Methodology	23
2.1 Introduction	24
2.2 Phases	25
2.3 Best practices for the publication of results	26

2.4	Evaluation	27
3	Publications	29
3.1	Definition of vocabularies and schemas	30
3.1.1	Onyx: Describing Emotions on the Web of Data	30
3.1.2	Onyx: A Linked Data Approach to Emotion Representation	43
3.1.3	Towards a Common Linked Data Model for Sentiment and Emotion Analysis	68
3.1.4	A Linked Data Model for Multimodal Sentiment and Emotion Analysis	76
3.1.5	EUROSENTIMENT: Linked Data Sentiment Analysis	86
3.1.6	Linguistic Linked Data for Sentiment Analysis	91
3.1.7	A Linked Data Approach to Sentiment and Emotion Analysis of Twitter in the Financial Domain	100
3.1.8	Generating Linked-Data based Domain-Specific Sentiment Lexicons from Legacy Language and Semantic Resources	113
3.2	Linked Data tools and services for sentiment analysis	118
3.2.1	Senpy: A framework for semantic sentiment and emotion analysis services	118
3.2.2	Multimodal Multimodel Emotion Analysis as Linked Data	125
3.2.3	MixedEmotions: An Open-Source Toolbox for Multi-Modal Emotion Analysis	132
3.2.4	Senpy: A Pragmatic Linked Sentiment Analysis Framework	145
3.3	Social context for Sentiment Analysis	154
3.3.1	Social context in sentiment analysis: Formal definition, overview of current trends and framework for comparison	154
3.3.2	A Model of Radicalization Growth using Agent-based Social Simulation	193
3.3.3	Modeling Social Influence in Social Networks with SOIL, a Python Agent-Based Social Simulator	210

3.3.4	Soil: An Agent-Based Social Simulator in Python for Modelling and Simulation of Social Networks	216
4	General Discussion, Conclusions and Future Research	229
4.1	Overview	230
4.2	Scientific results	232
4.2.1	Objective 1: Definition of a vocabulary for emotions	233
4.2.2	Objective 2: Definition of a model to annotate language resources and to be used in analysis services	235
4.2.3	Objective 3: Definition of a reference architecture for sentiment and emotion analysis services	241
4.2.4	Objective 4: Development of a reference implementation of the architecture	242
4.2.5	Objective 5: Modelling the types of contextual information and social theories	244
4.2.6	Other	247
4.3	Applications	248
4.4	Conclusions	253
4.5	Future Research	255
	Bibliography	259
	List of Figures	269
	List of Tables	271
	Glossary	273
	Appendix A Publications	275
A.1	Summary of publications	275
A.2	Publications indirectly related to the thesis	279

A.2.1	A Big Linked Data Toolkit for Social Media Analysis and Visualization based on W3C Web Components	279
A.2.2	An Emotion Aware Task Automation Architecture Based on Semantic Technologies for Smart Offices	298
A.2.3	Enhancing Deep Learning Sentiment Analysis with Ensemble Tech- niques in Social Applications	319
A.2.4	A modular architecture for intelligent agents in the evented web	331
A.2.5	Applying Recurrent Neural Networks to Sentiment Analysis of Spanish Tweets	332
A.2.6	Aspect based Sentiment Analysis of Spanish Tweets	339
A.2.7	MAIA: An Event-based Modular Architecture for Intelligent Agents .	346
A.2.8	EuroLoveMap: Confronting feelings from News	355

Reader's guide

This thesis is presented in the form of a compendium of publications (*compendio de publicaciones*). This means that this document is not a monograph, its core is a selection of publications by the author, which have been published in peer-reviewed journals and conferences and should support the quality and relevance of the results of this thesis. In particular, at least three of these publications must be published in journals in Q1 or Q2 positions of the JCR or Scopus ranking, and have the PhD student as first author. These requirements can be verified through list of publications in Annex A.

This form of thesis also imposes some additional requirements on the structure and content of the document. This section has been added to aid the reader navigate the document and evaluate its content. The document is structured as follows.

Chapter 1 is an introduction to this thesis, it explains the motivation, background and other relevant aspects. It also details the expected objectives and the initial hypotheses. The last section discusses how these objectives were met, and how the hypotheses were supported.

Chapter 2 explains the methodology used to conduct and evaluate the results of this thesis. It presents the phases in which research has been conducted. It also introduces the best practices followed for each type of result: definition of vocabularies, software, and other academic results. Lastly, it explains the evaluation criteria for each type of result.

Chapter 3 contains the full texts of novel publications that are core to this thesis. Each paper is presented in a separate section, where the full text is preceded by a short table with relevant information (title, authors, etc.), Publications are grouped into three categories, based on their topic: definition of vocabularies vocabularies and schemas, Linked Data for sentiment analysis, and social context for sentiment analysis. There are several tables of publications throughout the document, and all of them refer to the paper's section in this chapter. Some publications by the author have not been included in this chapter, as they do not directly contribute to the objectives. A complete list of relevant publications can be found in Annex A.

Chapter 4 is a general discussion about this thesis. It presents an overview of the

solutions proposed, analyzes the results, summarizes our conclusions from this work, and discusses future lines of research. The overview section (4.1) is aimed to connect all the contributions, and to help contextualize them. The results are presented in Section 4.2, and they are grouped by the objective (Section 1.4) they helped achieve. As a whole, this chapter is required to describe how each publication in Chapter 3 has contributed to this thesis, and their role to fulfil the objectives defined in Section 1.4. It should also justify the unity and coherence of the solutions and results.

Lastly, a summary of all publications is available in Annex A, including those publications that are indirectly related to this thesis.

CHAPTER 1

Introduction

This chapter presents the context of this PhD thesis. This includes its motivation, relevant technical and social aspects, as well as its research objectives.

1.1 Motivation

With the rise of social media, more and more users are sharing their opinions and emotions online (Pang and Lee, 2008a). The increasing amount of information and number of users is drawing the attention of researchers and companies alike, which seek not only academical results but also profitable applications such as brand monitoring. As a result, many tools and services have been created to enrich or make sense of human generated content. Extracting sentiment from text is a branch of NLP known as sentiment analysis.

There are some terminological issues with terms such as “sentiment analysis” and “opinion mining”, which are sometimes used interchangeably. This has been a long standing debate that predates the creation of online social networks. (Pang and Lee, 2008a) explains the evolution of each of these terms, and how they have been used in the literature. According to them, when broad interpretations are applied, sentiment analysis and opinion mining denote the same field of study (which itself can be considered a sub-area of subjectivity analysis). For our purposes, we will use the term sentiment in its broadest sense, and the term affect as an umbrella in the rare occasions where the distinction between sentiment and opinion is important. We will also use the term sentiment analysis services, which are services that provide sentiment analysis of some sort, often over an HTTP protocol.

Terminological issues aside, sentiment analysis is broadly understood as the classification of text as positive or negative. In other words, polarity classification. Some works also introduce a “neutral” label, or finer grained categories such “very positive” or “mildly negative”. Some other works use a continuous scale, and content is given a polarity value, often in the -1 to 1 range, although this varies from work to work. All these are variations over the same type of analysis.

This thesis focuses on different ways in which sentiment analysis can be improved, from incorporating new concepts and tools to improve usability, to leveraging sources other than text. As illustrated Figure 1.1, we will cover four types of improvement: the addition of Linked Data to represent both language resources and analysis services, the move from sentiment analysis to emotion analysis, the combination of analysis in text with other types of analysis (multi-modal analysis), and how more information about the social network can be leveraged.

It is important to note that these changes can be added independently, i.e., an analysis can use social context without dealing with emotions or other modalities, but it may as well use all the elements at the same time. We will see what each of these changes mean, why they are important for the future of sentiment analysis, and the challenges they impose.

Figure 1.1: Overview of the evolution of sentiment analysis, from pure text to using information.

So far sentiment analysis tools and services have used ad-hoc annotation schemata and services remain as isolated data silos. The lack of consensus is detrimental in several ways:

1. it impedes the creation of a homogeneous ecosystem of tools, libraries and applications;
2. it introduces ambiguity in the annotation (e.g., some services use a polarity scale of 0 to 1, whereas others use -1 to 1);
3. it hinders user adoption;
4. it makes evaluation of different services and tools hard.

These are strong barriers for the progress of the field.

Moreover, as more resources and applications appear, new types of sentiment analysis that account for more complex types of affects have started to emerge. For instance, some

works focus on classifying the emotion shown by a user in a specific text, such as happiness or surprise. Other works do stance classification (Pamungkas, Basile, and Patti, 2019), or detecting whether a person supports, denies, queries, or comments on a specific issue. These types of analysis can be seen as more complex forms of sentiment or subjectivity analysis, which pose whole new set of challenges.

In particular, in this thesis we focus on emotion analysis. Emotions have a crucial role in our lives, and even change the way we communicate (Pang and Lee, 2008a). They can be passed on just like any other kind of information, in what some authors call emotional contagion (Barsade, 2002). That is a phenomenon that is clearly visible in social networks (Kramer, Guillory, and Hancock, 2014), and it can be observed in Online Social Network (OSN) through public APIs make it relatively easy to study the social networks and their information flow.

Some social sites are already using emotions natively, giving their users the chance to share emotions or use them in queries. This is exemplified by Facebook, which recently updated the way its users can share personal statuses. And there are some tools and languages to annotate different types of media with emotions, such as EmotionML. However, there is no common semantic vocabulary for emotions.

On the other hand, natural language is often mixed with other types of information that can be analyzed. For instance, a video of a speech has at least three different elements that can be analyzed using different techniques: 1) the script or words can be analyzed using NLP; 2) the voice is analyzed using audio analysis; 3) the posture and gestures can be analyzed using video analysis. The combination of all these analyses is known as multi-modal analysis, and it is an active field that gathers experts from different disciplines. Each type media lends itself to different types of analysis. In fact, one of the reasons emotion detection in text far less popular than polarity classification may be that it is easier to detect the valence (positive or negative) of an opinion in text, than it is to detect other dimensions of an emotion, such as the arousal (high intensity level) of the author.

Emotion analysis is more common in audio than in text, as the opposite occurs: it is harder to detect valence than arousal or dominance. Combining different modalities can help counter the weaknesses of each modality, but it can be very challenging, as each field has traditionally worked in isolation, using different tools, naming conventions and models. This leads to technical challenges, such as dealing with different tools and formats, and some conceptual challenges, such as having different emotion models. The technical challenges are mostly caused by the lack of communication between communities, and they can be remedied through tooling and using linked data vocabularies as lingua franca. Conceptual

differences such as model heterogeneity are trickier to solve. Some of these differences stem from the lack of a standard affect model, and the fact that certain models are more suited to specific modalities. As we said, it is easier to detect arousal (high intensity levels) from audio than it is to detect valence (polarity). This may be a reason why the speech analysis community has used dimensional models such as the Valence-Arousal-Dominance model, whereas the text community has relied on positive, negative and neutral tags. Some types of analysis may require converting all annotations to a common model, which is not trivial.

Lastly, despite the advantages of multi-modal analysis, simply analysing the content of a message is still not enough to confidently infer the sentiment of the author. When humans communicate, there is a plethora of information that is assumed or implicitly known, such as known political affiliations of each interlocutor, history of comments, or acquaintances. Social media content relies heavily on this type of information, due to its informal tone and the brevity of the interactions. If this information could be modelled and exploited, it may improve sentiment analysis in social media, as shown by some related works that have already shown promising results.

We have thus identified four shortcomings of the state of the art in sentiment analysis in text: 1) lack of interoperability, i.e., heterogeneity of formats and schemas; 2) underrepresentation of emotion analysis; 3) difficulty to integrate with other types of analysis; and 4) disregard for contextual information. To target interoperability, Linked Data can be used as a lingua franca for data representation as well as a set of tools to process and share such information. Plenty of services have embraced the Linked Data concepts and are providing tools to interconnect the previously closed silos of information (Tummarello, Delbru, and Oren, 2007). It has even proven useful in some areas of Opinion Mining. Some schemata offer semantic representation of opinions (Westerski, Carlos A. Iglesias, and Tapia, 2011), allowing richer processing and interoperability. The same could be applied to sentiment analysis as a whole, and to Emotion Analysis, once there is a proper model for emotions. In addition to vocabularies and schemas, achieving real service interoperability would further require common APIs and methodologies. Once a proper model for emotions is in place, new resources and examples for emotion analysis should become available, thus increasing the popularity of emotion analysis. A Linked Data principled approach would also help with multimodal analysis by providing common models for all modalities, modelling all the transformations necessary to fuse different modalities, and helping track provenance information.

In summary, sentiment and emotion analysis could benefit of the combination of a linked data principled approach, multi-modal analysis, and leveraging contextual information. The

first one has the potential to enable novel applications, interoperability of tools and services, and in general would foster research and applications in the field. The last two may be used to improve the performance of sentiment and emotion classification. This thesis aims to tackle these issues, effectively modelling and linking affects (sentiment and emotion), data (knowledge graph), and people (context in the social network).

1.2 Background

1.2.1 Sentiment and Emotion Analysis

Sentiment analysis has been an active field of research for a long time, but it has grown in popularity with the advent of online opinion-rich resources (Pang and Lee, 2008b). This new type of content comes with new challenges and limitations. The quintessential example is how microblogging sites such as Twitter impose hard limitations on text extension. In the case of Twitter, that limitation has only been recently lifted from the iconic 140 characters. Due to the limitations and the intended audience, those texts also tend to be more relaxed and make heavy use of slang and abbreviations. In contrast, prior to the creation of these OSN, sentiment analysis dealt mostly with longer and formal texts, such as those in news articles, papers and books. A great deal of the research in recent years has gone into finding ways to combat those challenges. In turn, these resources have also added their own set of limitations and challenges.

Over the last two decades, numerous works have explored different approaches to sentiment analysis. These approaches can be grouped into machine learning, lexicon based, and hybrid (K. Ravi and V. Ravi, 2015). Of the three, machine learning techniques and hybrid approaches seem to be dominant (Araque et al., 2017; Pang, Lee, and Vaithyanathan, 2002; Wang and Manning, 2012), and lexicon techniques are typically incorporated into machine learning approaches to improve their results. Machine learning approaches apply a predictor (a classifier, or an estimator) on a set of features that represent the input. The set of predictors is not very different from those used in other areas. Instead, the complexity in these approaches lies in extracting complex features from the text, filtering only relevant features, and selecting a good predictor (Sharma and Dey, 2012).

One of the most straightforward features is the Bag Of Words (BOW) model. In BOW, each document is represented by the multiset (bag) of its constituent words. Word order is disrupted, and syntactic structures are broken. As a result, a great deal of information from natural language is lost (Xia and Zong, 2010). Therefore, various types of features have been exploited, such as higher order n-grams (Pak and Paroubek, 2010). A more sophisticated

feature is Part of Speech (POS) tagging (Gimpel et al., 2011). In it, a syntactic analysis process is run, and each word is labeled (tagged) with its syntactic function (e.g., noun). Additionally, syntactic trees can be calculated. Using these trees, the words in the input can be rearranged to a more convenient position while still conveying the same meaning. Note how these two types of features only rely on lexical and syntactical information. For this reason, they are sometimes referred to as surface forms.

Surface forms can also be combined with other prior information, such as word sentiment polarity (García-Pablos, Cuadros Oller, and Rigau Claramunt, 2016; Cambria, 2016; Kiritchenko, Zhu, and Mohammad, 2014; Melville, Gryc, and Lawrence, 2009; Nasukawa and Yi, 2003). This prior knowledge usually takes the form of sentiment lexicons, i.e., dictionaries that associate words in a domain or language with a sentiment. Some lexicons also include non-words such as emoticons (Jiang et al., 2015; Hogenboom et al., 2015) and emoji (Novak et al., 2015). These alternative forms of writing have been shown very useful, as they can dominate textual cues and form a good proxy for text polarity (Hogenboom et al., 2015).

The use of lexicon-based techniques has many advantages (Taboada et al., 2011), most of which stem from their combination with other methods. For instance, it is possible to generate lexicons that are domain dependent or that incorporate language-dependent characteristics. Lexicons and syntactic information can also be combined with linguistic context to shift valence (Polanyi and Zaenen, 2006). On the other hand, there are several disadvantages to lexicon approaches. First, creating lexicons is an arduous task, as it needs to be consistent and reliable (Taboada et al., 2011). It also needs to account for valence variability across domains, contexts, and languages. These dependencies make it hard to maintain domain-independent lexicons. An alternative to retain independence while encoding domain, language, and context variability is through semantic representation of the lexical resources in the form of ontologies. An ontology can encode both lexical (McCrae, Spohr, and Cimiano, 2011b) and affective (Sánchez-Rada and Carlos A. Iglesias, 2016) nuances, both in the lexicons and in the automatic annotations (Buitelaar, Arcan, et al., 2013). This is especially useful for aspect-based sentiment analysis, as the differences between aspects can be incorporated into the ontology (Wei and Gulla, 2010).

In recent years, new approaches based on deep learning have shown excellent performance in Sentiment Analysis (Collobert et al., 2011; Bengio, 2009). In contrast with traditional techniques, deep learning techniques learn complex features from data with minimum human interaction. These algorithms do not need to be passed manually crafted features: they automatically learn new complex features. The downside is that the quality of the features

heavily depends on the size of the training data set. Hence, they often require large amounts of data, which is not always available. They also raise other concerns such as interpretability (Marcus, 2018; Lipton, 2016) or its inability to adapt to deal with edge cases (Marcus, 2018). In the realm of NLP, most of the focus is on learning fixed-length word vector representations using neural language models (Kim, 2014). These representations, also known as word embeddings, can then be fed into a deep learning classifier, or used with more traditional methods. One of the most popular approaches in this area is word2vec (Mikolov et al., 2013). The downside of these methods is that they require enormous amounts of training data. Luckily, several researchers have already applied these methods to large corpora such as Wikipedia and released the resulting embeddings.

Lastly, it is also possible to combine independent predictors to achieve a more accurate and reliable model than any of the predictors on their own. This approach is known as ensemble learning. Many ensemble methods have been previously used for sentiment analysis. Ensemble methods can be classified according to two main dimensions Rokach (2010): how predictions are combined (rule-based and meta-learning), and how the learning process is done (concurrent and sequential). A new application of ensemble methods is the combination of traditional classifiers based on feature selection and deep learning approaches (Araque et al., 2017).

1.2.2 Linked Data for NLP

Tim Berners-Lee, 2006 outlined a set of rules for publishing data on the Web in a way that all published data becomes part of a single global data space:

1. Use URIs as names for things
2. Use HTTP URIs so that people can look up those names
3. When someone looks up a URI, provide useful information, using the standards(RDF, SPARQL)
4. Include links to other URIs, so that they can discover more things

In Berner-Lee's own words, rather than rules these are actually expectations of behavior: breaking them does not destroy or break any contract. But by following them we seize the opportunity to data interconnected. Hence, these rules have since been known as the Linked Data principles In general, Linked Data Bizer, Heath, and Berners-Lee, 2009 refers to best practices and technologies for publishing, sharing and connecting structured data on the web.

What is missing from those rules is how the useful information should be organized. It leaves that side to Resource Description Framework (RDF) and other standards. RDF is a standard for establishing semantic interoperability on the Web Decker et al., 2000. It is similar to XML in many aspects. Except that that XML only addresses document structure, i.e., the hierarchical relation of different elements, and their attributes. By contrast, RDF provides a data model that can be extended to provide more sophisticated ontological representations. This flexibility means that RDF better facilitates interoperability.

At its core, the RDF model is very simple. It is based on subject-predicate-object expressions. For instance, let us consider the the expression “I studied at UPM”. Here, the subject is “I”, the predicate is “studied at”, and the object is “UPM”. For all three, we would need a URI, or unique identifier. As their name implies, these URIs must be unique to each specific concept, and they should be addressable to comply with the linked data principles. We will use an invalid domain as base, and the following full URIs: `http://example.com/jfernando`, `http://example.com/studied-at` and, lastly `http://example.com/UPM`. By the same token, we may add new predicates and URIs. For instance, the example in Listing 1.1 encodes the information: my supervisor and I studied at UPM.

Listing 1.1: Example of RDF in n-triples notation

```
<http://example.com/jfernando> <http://example.com/studied-at> <http://example.com/UPM> .  
<http://example.com/jfernando> <http://academia.test/supervised-by> <http://example.com/cif> .  
<http://example.com/cif> <http://example.com/studied-at> <http://example.com/UPM> .
```

Writing all triples explicitly is very cumbersome and it reduces readability. RDF has other representation formats that are better for human consumption, such as Turtle (Listing 1.2). These formats can also use additional features such as prefixes, to make the annotation even easier to read and write.

Listing 1.2: Example of RDF in turtle notation

```
@prefix ex: <http://example.com/> .  
@prefix academia: <http://academia.test/> .  
  
ex:jfernando ex:studied-at ex:UPM ;  
             academia:supervised-by cif.  
  
cif ex:studied-at ex:UPM .
```

An interesting alternative format is JSON-LD¹, a subset of JSON that incorporates se-

¹<http://json-ld.org>

mentics. JSON-LD was designed as a lightweight Linked Data format for human consumption and creation. With other RDF representation formats, new tools, standards (SPARQL) and software (e.g., databases) are required to store and query data. Since JSON-LD is fully compatible with all the JSON ecosystem, it can be used anywhere JSON is. This makes it an ideal data format for programming environments, REST Web services, and unstructured databases such as CouchDB and MongoDB. The previous examples could be represented in JSON-LD as Listing 1.3. The semantics of each property and value are provided by the JSON-LD context, which can be re-used and referenced by URL, such as in the example. In our example, the context is shown in Listing 1.4.

In contrast with the graph-based RDF model, JSON uses a hierarchical model-based structure. This means that there are different ways (i.e., schemas) to represent an RDF graph in JSON-LD form. For each situation, some of those schemas are more verbose, and some are easier to use. Part of the challenges in defining a model for sentiment and emotion analysis will lie in selecting a schema that is ergonomic while retaining all the necessary information.

Listing 1.3: Example of RDF in JSON-LD notation

```
{ "@context": "http://example.com/context.jsonld",
  "@id": "jfernando",
  "studied-at": "UPM",
  "supervised-by": {
    "@id": "cif",
    "studied-at": "UPM"
  }
}
```

Listing 1.4: Context for Listing 1.3

```
{
  "@context": {
    "@base": "http://example.com",
    "supervised-by": {
      "@id": "http://academia.test/supervised-by"
    },
    "studied-at": {
      "@type": "@id",
      "@id": "http://example.com/studied-at"
    }
  }
}
```

```
}
```

So far, we have defined our own URIs and predicates. In other words, we have chosen our representation model for our example domain. This is fine for an ad-hoc case, but one of the advertised advantages of Linked Data is the ability to connect different data sources. If different data sources use very different predicates and URIs for the same concepts, we will not be able to merge their information without intervention. Consequently, we need a way to agree on a set of common predicates and entities that are valid. These models are also referred to as ontologies, vocabularies, or specifications. There are different types of vocabularies. The simplest are just a list of predicates and entities. Some are expressed in rich languages such as Web Ontology Language (OWL) and allow for a fine-grain definition of the domain, which can later be used for reasoning and inference. These ontologies add useful concepts such as classes, subclasses, properties, ranges and domains. Back to our example, we may add classes to our vocabulary such as `PhD student`, and properties such as `date of defense`. We may also include restrictions like *people can only study at a university*, or *people can only be supervised by full professors*. This could be used to check for incoherent data (e.g., if `ex:cif` was listed as a fellow student (not a professor)), or to infer missing data.

To foster reusability and composability, each vocabulary is expected to cover one domain, and to rely on others to model aspects outside of that domain. For instance, in our example we could add my e-mail address re-using properties from other vocabularies such as Friend of a Friend (FOAF). An example of vocabulary that is very widespread on the internet is `schema.org`², which covers different domains that are useful from a web search engine, such as venues, online reviews and people.

This thesis covers three different domains that need to be modelled with linked data: 1) language resources 2) sentiment and emotion analysis services. Rather than creating an ad-hoc model for each domain, linked data principles encourage reusing already existing models; 3) sentiment and emotions (more generally, affects);.

Linked data technologies have gained wide acceptance in the realm of language resources, as a mechanism to combine heterogeneous resources. Before the use of linked data, several initiatives had addressed interoperability of language resources since the late 1980s such as Text Encoding Initiative (TEI) (Ide and Veronis, 1995), but there was not yet a widely accepted global solution for integrating and combining heterogeneous linguistic resources from different sources (Chiarcos, 2013). In particular, the Linked Open Data (LOD) Project

²<http://schema.org>

is a grassroots community effort supported by W3C whose aim is to bootstrap the Web of Data by identifying existing datasets available under open licenses, convert them to RDF following the Linked Data principles, and to publish them on the Web. The data cloud originated from this initiative is known as the LOD Cloud. Several communities such as Open Linguistics Working Group (OWLG) (Chiarcos, 2013) proposed the idea of adopting linked data principles for representing, sharing and publishing open linguistic resources with the aim of developing a sub-cloud of LOD cloud of linguistic resources, known as the Linguistic Linked Open Data (LLOD) cloud (Chiarcos, Hellmann, and Nordhoff, 2011a).

In addition, the use of linked data for modeling linguistic resources provides a clear path to their semantic annotation and linking with semantic resources of the Web of Data. This is especially important for making sense of social media streams whose semantic interpretation is particularly challenging, because they are strongly inter-connected, temporal, noisy, short, and full of slang (Bontcheva and Rout, 2012). Moreover, several authors (Saif, He, and Alani, 2012) have shown that the use of semantics in sentiment analysis outperforms semantics-free methods. Thus, the availability of semantically annotated linguistic resources is a crucial to the development of the field of sentiment analysis.

Similarly, the NLP community has recently started to use alternatives to model NLP processes as linked data, and to achieve interoperability between different NLP tools and services. The most popular initiative in this area is NLP Interchange Format (NIF). NIF 2.0 (Hellmann et al., 2013) defines a semantic format and API for improving interoperability among natural language processing services. NIF follows a linked data principled approach so that different tools or services can annotate a text. To this end, texts are converted to RDF literals and an URI is generated so that annotations can be defined for that text in a linked data way. NIF offers different URI Schemes to identify text fragments inside a string, e.g. a scheme based on RFC5147 (Wilde and Duerst, 2008), and a custom scheme based on context.

Thus, NIF is a natural fit for sentiment and emotion analysis in text. But, for that, NIF needs to be extended to include specific parts of Sentiment and Emotion Analysis, such as a description of Opinions and Emotions. That is the role of vocabularies such as Marl.

Marl (Westerski, Carlos A. Iglesias, and Tapia, 2011) is a vocabulary for Opinion Mining. It was designed to annotate and describe subjective opinions expressed in text. In essence, it provides the conceptual tools to annotate Opinions and results from Sentiment Analysis in an open and sensible format. However, it is focused on polarity extraction and is not capable of representing Emotions. Onyx aims to remedy this and offer a complete set of tools for any kind of Sentiment Analysis, including advanced Emotion Analysis.

LLOD (Chiarcos, Hellmann, and Nordhoff, 2011b; Chiarcos, McCrae, et al., 2013) is an initiative that promotes the use of linked data technologies for modeling, publishing and interlinking linguistic resources. The main benefit of using linked data principles to model linguistic resources is that it provides a graph-based model that allows representing different kinds of linguistic resources (such as lexical-semantic resources, linguistic annotations or corpora) in a uniform way, thus supporting querying across resources. The creation of an LLOD cloud is a cooperative task that is been managed by several communities, such as OWLG of the Open Knowledge Foundation and Ontology-Lexica Community Group (OntoLex) of W3C. As a result of this activity, an initial LLOD is currently available, where several types of resources have been identified: lexical-semantic resources (e.g. machine readable dictionaries, semantic networks, semantic knowledge bases, ontologies and terminologies), annotated corpora and linguistic annotations.

With regards to modeling lexical-semantic resources, we find *Lexicon Model for Ontologies (lemon)* (Buitelaar, Cimiano, et al., 2011). lemon proposes a framework for modeling and publishing lexicon and machine-readable dictionaries as linked data. lemon provides a bridge between the most influential lexical-semantic resources, WordNet (Fellbaum, 1998) and DBPedia (Bizer, Lehmann, et al., 2009). Lemon was designed to meet the following challenges:

- RDF-native form to enable leverage of existing Semantic Web technologies (SPARQL, OWL, RIF etc.).
- Linguistically sound structure based on LMF to enable conversion to existing offline formats.
- Separation of the lexicon and ontology layers, to ensure compatibility with existing OWL models.
- Linking to data categories, in order to allow for arbitrarily complex linguistic description. In particular the LexInfo vocabulary is aligned to lemon and ISOcat.
- A small model using the principle of least power - the less expressive the language, the more reusable the data.

With regards to annotated corpora, there are two initiatives (Chiarcos, Hellmann, and Nordhoff, 2012), POWLA (Chiarcos, 2012b) and NIF (Hellmann, 2013) that enable to link lexical-semantic resources to corpora. Finally, Ontologies of Linguistic Annotation (OLiA) (Chiarcos, 2012a) are a repository of annotation terminology for various linguistic phenomena that can be used in combination with POWLA, NIF or *lemon*. OLiA ontologies

allow to represent linguistic annotations in corpora, grammatical specifications in dictionaries, and their respective meaning within the LLOD cloud in an operable way.

The main benefits of modeling linguistic resources as linked data include (Chiarcos, McCrae, et al., 2013) interoperability and integration of linguistic resources, unambiguous identification of elements of linguistic description, unambiguous links between different resources, possibility to annotate and query across distributed resources and availability of mature technological infrastructure.

Another key aspect of the semantic models in this thesis is provenance. Provenance is information about entities, activities, and people involved in producing a piece of data or thing, which can be used to form assessments about its quality, reliability or trustworthiness. The PROV Family of Documents defines a model, corresponding serializations and other supporting definitions to enable the inter-operable interchange of provenance information in heterogeneous environments such as the Web (Moreau et al., 2011). It includes a full-fledged ontology that other ontologies can link to. The complete ontology is covered by the PROV-O Specification (Groth and Moreau, 2013). In essence, Agents take part in Activities to transform Entities (data) into different Entities (modified data). This process can be aggregation of information, translation, adaptation, etc. For the purpose of this thesis, this activity is usually a sentiment or emotion analysis, which turns plain data into semantic sentiment or emotion information. There are many advantages to adding provenance information in emotion analysis, as different algorithms may produce very different results that can then be compared or aggregated.

1.2.3 Emotion Representation

Emotion is a very common concept that gets used often in everyday life, it is ingrained in our vocabulary. The term usually gets used in a very loose and subjective way. For emotion analysis, and any other kind of research on affective computing, we need a more rigorous definition of emotion, including what emotions are possible, and how they can be measured and represented. We refer to that definition as a model of emotions, or emotion model. As it turns out, modelling emotions is a rather complex task. In this section we will cover the most popular attempts to propose a universal model of emotions. We will also cover some alternative ways to represent emotions in a machine-readable way.

Various representation schemes for emotions have been proposed over the years, each based on particular criteria, ranging from the most simplistic and ancient that come from Chinese philosophers to the most modern theories that refine and expand older models (Ekman, 1999; Prinz, 2004). The literature on the topic is vast, and it is out of the scope of

this section to cover all these models. Other works have presented a detailed overview of different emotion models and their use in affective computing (Cambria, Livingstone, and Hussain, 2012). Overall, there are two main types of models: models based on categories, and dimensional models. In the former, emotions are represented with one or more categories or labels (e.g., happy, sad). In practice, in addition to categories there may also be other numeral values, such as value (i.e., how strong the emotion is), and confidence of the annotation. On the other hand, dimensional models represent emotions in the continuum of an n -dimensional space. The number of dimensions varies from model to model. Additionally, some models label specific regions of that space.

One of the most popular categorical models, even in popular culture, is Ekman’s model of *six basic emotions* (Anger, Fear, Surprise, Happiness, Disgust, Sadness) are based on the universality of those emotions (Ekman and Friesen, 1971). The advantage of Ekman’s model is its simplicity. At the same time, other researchers have found this simplicity too limiting to represent the wealth. For instance, Plutchik’s *wheel of emotion* is an alternative to Ekman’s model that provides more categories of emotion. It is based on contrast and closeness of emotions (Plutchik, 1980a), and the categories are arranged in concentric circles of growing complexity. At its root, there are eight basic emotions: anger, fear, sadness, disgust, surprise, anticipation, trust, and joy. All other emotions are derived from those. Plutchik’s model has been extensively used (Borth et al., 2013; Cambria, Havasi, and Hussain, 2012) in the area of Sentiment Analysis and Affective Computing, relating all the different emotions to each other in what is called the rose of emotions. Other categorical models cover affects in general, which include Emotions as part of them. One of them is the work by Strapparava and Valitutti, WordNet-Affect (Strapparava and Valitutti, 2004). It comprises more than 300 affects, many of which are classified as emotions. What makes this categorisation interesting is that it effectively provides a taxonomy of emotions. It both gives information about relationship between emotions and makes it possible to decide the level of granularity of the emotions expressed. Lastly, the recent work by Cambria et al. (Cambria, Livingstone, and Hussain, 2012) introduces a novel model, The Hourglass of Emotions, inspired by Plutchik’s studies (Plutchik, 1980b).

On the dimensional front, we will cover three closely-related models. First, Russel’s *Circumplex model* is constructed to capture the core affect in a two dimensional (Arousal and Valence) model (Russell, 2003; Posner, Russell, and Peterson, 2005). Arousal reflects the level of energy in the emotion (e.g., pleased vs. ecstatic), whereas valence reflects the hedonic tone (e.g., pleasant vs. unpleasant). Osgood later identified an additional dimension, and the result is the very widespread PAD (Pleasure, Arousal, Dominance) three-dimensional representation (Osgood, Suci, and Tannenbaum, 1957). Dominance represents the sense of

control or dominant nature of the emotion (e.g., fear vs. anger). More recently, Fontaine et al. identified a fourth dimension (unpredictability) (Fontaine et al., 2007). This new dimension refers to the appraisal of expectedness or familiarity.

Despite the efforts put into these models, there does not seem to be a universally accepted model for emotions (Schröder, Pirker, and Lamolle, 2006). This poses a problem for any field of emotion research, but it is even more troubling for affective computing, where systems need a representation format for emotions. Emotion Markup Language (EmotionML) (Burkhardt et al., 2013) is one of the most notable general-purpose emotion annotation and representation languages. It was born from the efforts made for Emotion Annotation and Representation Language (EARL) (Excellence, 2006; Schröder, Pirker, and Lamolle, 2006) by Human-Machine Interaction Network on Emotion (HUMAINE).

In a discussion regarding EmotionML, Schröder et al. pose that *any attempt to propose a standard way of representing emotions for technological contexts seems doomed to fail* (Schröder, Devillers, et al., 2007). Instead they claim that the markup should offer users choice of representation, including the option to specify the affective state that is being labelled, different emotional dimensions and appraisal scales. The level of intensity completes their definition of an affect in their proposal. As a result, acEmotionML offers twelve vocabularies for categories, appraisals, dimensions and action tendencies. A vocabulary is a set of possible values for any given attribute of the emotion. There is a complete description of those vocabularies and its computer-readable form available (Ashimura et al., 2012).

There have also been some efforts to provide a semantic vocabulary for emotions. First, the Chinese Emotion Ontology (Yan et al., 2008) was developed to help understand, classify and recognize emotions in Chinese. The ontology is based on HowNet, the Chinese equivalent of WordNet. The ontology provides 113 categories of emotions, which resemble the WordNet taxonomy and the authors also relate the resulting ontology with other emotion categories. All the categories together contains over 5000 Chinese verbs. Soon after, Grassi presented Human Emotion Ontology (HEO) (Grassi, 2009). This ontology presents an ontology for human emotions for its use for annotating emotions in multimedia data. HEO deals with the heterogeneity of emotional theories by providing a generic characterisation of emotions. There are several limitations with HEO. First of all, despite being a generic ontology, HEO embeds the most common models (Plutchik, Ekman, Hourglass, etc.) in the ontology itself. Moreover, its representation of action tendencies and appraisal as entities, instead of properties, makes its use for dimensional annotation very verbose. It also includes some properties (e.g. *isAnnotatedBy*) that, while useful for certain types of

Emotion Analysis, are not generic or there are other established ontologies that provide the same concept. And it makes assumptions about certain properties, which limits its use. For instance, automated emotion analysis would not have a human annotator, as HEO suggests.

Another work worth mentioning is that of Hastings et al. (Hastings et al., 2011) in Emotion Ontology (EMO), an ontology that tries to reconcile the discrepancies in affective phenomena terminology. It is, however, too general to be used in the context of emotion analysis: it provides a qualitative notion of emotions, when a quantitative one would be needed.

1.2.4 Contextual information for sentiment analysis

Contextual information refers to the collection of users, content, relations, and interactions which describe the environment in which social activity takes place. It encapsulates the frame in which communication in social media takes place. Since this is a new field of research, the definition and use of this contextual information for sentiment analysis was rather vague. This makes it hard to describe or compare works of the state of the art. One of the contributions of this thesis is a formal definition of social context. For the sake of clarity, we will use that terminology throughout the thesis.

Social context is used in sentiment analysis for two reasons that are subtly different. First, it can be used to compensate for implicit elements in the text. An example of this is how slang, abbreviations or semantic variations can be detected and accounted for in the classification. Humans apply a similar process when trying to understand content. Content authors also unconsciously rely on this fact and they assume certain prior knowledge. The second motivation to add social context is that it may help correct ambiguity or situations where textual queues are lacking. For example, a classifier may use the sentiment of earlier posts by the user and similar users on the same topic.

Tan et al. (2011) is one of the first works to incorporate social context information, which the authors called heterogeneous graph on topic, to infer (user) sentiment. The underlying ideas behind that work are user consistency and homophily. A function to measure each of those attributes is provided, and the model tries to maximize the overall value. The authors compare alternative ways to construct the user network, using variations of follower-followee relations and direct replies (interactions). In their results, relations and interactions yield similar results. In the original formulation edges (relations or interactions) are not weighted, so users are influenced equally by all their neighbors. Interactions are bound to be noisy, and aggregating them in this fashion is likely to provide little or no advantage over a simple relation. The SANT model (X. Hu et al., 2013) follows similar ideas but

for content classification. It also combines sentiment consistency, emotion contagion and a unigram model in a classifier.

Pozzi et al. (2013) extended the model by Tan et al. (2011). Their model uses what they call an approval network, which effectively add weights for edges between users. The rationale for that change is that friendship does not imply approval, and that a weighted network of interactions should better capture emotion contagion.

Other models have also exploited other strategies such as community detection in their analysis. An example is Xiaomei et al. (2018), which incorporate weak dependencies between microblogs, using community detection (different algorithms) on a network of microblogs. In their work, microblogs are connected if their authors are (i.e., there is a follower-followee relation). They refer to connections inferred from community detection as *weak connections*.

1.2.5 Social Network Analysis and Community Detection

Social Network Analysis (SNA) is the investigation of social structures Otte and Rousseau, 2002. It provides techniques to characterize and study the connections between people, and their interactions. SNA is not limited to OSN, but to any kind of social structure. Other examples of social network would be a network of citations in publications or a network of relatives. Through SNA techniques, it is possible to extract information from a social network that may be useful for sentiment analysis, such as chains of influence between users, groups of like-minded users, or metrics of user importance.

There are several ways in which SNA techniques can be exploited in sentiment analysis, but most of them fall under one of two categories: those that transform the network into metrics or features that can be used to inform a classifier; and those that limit the analysis to certain groups or partitions of the network.

A simple example of metrics provided by SNA could be user’s follower in-degree (number of users that follow the user) and out-degree (number of users followed by the user), which could be used as features for each user (Sixto, Almeida, and López-de-Ipiña, 2018). However, these metrics are not very rich, as they only cover users directly connected to a user, and it does so in a very naive way: all connections are treated equally. Other more sophisticated metrics could be used instead of in/out-degree, such as centrality, a measure of the importance of a node within a network topology, or PageRank, an iterative algorithm that weights connections by the importance of the originating user. Several works have introduced alternative metrics for user and content influence in a network (Hajian and White, 2011; Noro and Tokuda, 2016).

The second category of approaches is what is known either as network partition or as community detection, depending on whether the groupings may overlap. Intuitively, community detection aims to find subgroups within a larger group. This grouping can be used to inform a classifier, or to limit the analysis to relevant groups only. More precisely, community detection identifies groups of vertices that are more densely connected to each other than to the rest of the network (Papadopoulos et al., 2012). The motivation is to reduce the network into smaller parts that still retain some of the features of the bigger network. These communities may be formed due to different factors, depending on the type of link used to connect users, and the technique used to detect the communities. Each definition has its own set of characteristics and shortcomings. For instance, if users are connected after messaging each other, community detection may reveal groups of users that communicate with each other often (Deitrick and W. Hu, 2013). By using friendship relations, community detection may also provide the groups of contacts of a user (Gao et al., 2012).

The reader is referred to other publications (Papadopoulos et al., 2012; Orman, Labatut, and Cherifi, 2011) for further details of the different definitions of community and algorithms to detect them.

1.3 Hypotheses

After taking into consideration the background and context of this thesis, we formulated the following hypotheses:

- **Hypothesis-1** A Linked Data approach to sentiment analysis would increase interoperability between services and enable advanced capabilities such as automatic evaluation
- **Hypothesis-2** A semantic vocabulary for emotions would ease multi-modal analysis and enable the use of different emotion models
- **Hypothesis-3** Sentiment of social media text can be predicted using additional contextual information (e.g., previous history and relations between users)

1.4 Objectives

In order to test the hypotheses in the previous section, we set the following objectives:

- **Objective-1** Definition of a vocabulary for emotions

The vocabulary should be applicable to different use cases, including language resources (lexica and corpora) and analysis services.

- **Objective-2** Definition of a model to annotate language resources and to be used in analysis services

The model should leverage existing vocabularies such as Marl and lemon, and provide a unified representation for both language resources and services.

- **Objective-3** Definition of a reference architecture for sentiment and emotion analysis services

The architecture will be based on the definition of a REST-ful resource model, in order to facilitate the development of services. The architecture should be adaptable to a Big Data scenario.

- **Objective-4** Development of a reference implementation of the architecture

The reference implementation will be valuable to evaluate the soundness of the architecture.

- **Objective-5** Modelling the types of contextual information and social theories (such as emotion propagation) that can be leveraged for sentiment and emotion analysis

This may require the use of additional models, such as social theories of behavior and group formation, or techniques from other fields, such as Social Network Analysis.

1.5 Document outline

The rest of the document is structured as follows: Chapter 2 explains the methodology used to conduct and evaluate the results of this thesis, the phases of this research, best practices followed and evaluation criteria for each type of result; Chapter 3 contains the full texts of novel publications that are core to this thesis; Each paper is presented in a separate section, where the full text is preceded by a short table with relevant information (title, authors, etc.); Publications are grouped into three categories, based on their topic: definition of vocabularies vocabularies and schemas, Linked Data for sentiment analysis, and social context for sentiment analysis; Chapter 4 is a general discussion about this thesis; It presents an overview of the solutions proposed, an analysis of the results, a summary of our conclusions and a discussion of future lines of research. Lastly, a summary of all publications is available in Annex A, including those publications that are indirectly related to this thesis.

CHAPTER 2

Methodology

In layman's terms, research is 'finding out something you don't know.' Luckily, the actual act of research is much less vague. It needs to be methodical, focused and rigorous. This chapter presents the methodological considerations to ensure this thesis follows those principles.

2.1 Introduction

There are two ways to conduct research, using induction or deduction. The first way proposes generalizations based on observations. For this camp, the cornerstone of research is the examination of the adequacy of generalizations, formulated as hypotheses (Phillips, Pugh, et al., 2010)¹. The deductive approach uses known theory to propose hypotheses, which are then tested. A proponent of this view was Karl Popper who claimed that the nature of scientific method is hypothetico-deductive and not, as is generally believed, inductive. When conducting a PhD thesis, it is important to understand the difference between these two interpretations of the research process to avoid discouragement, and suffering from a feeling of ‘cheating’ or not going about it the right way.

A popular misconception about scientific method is that it is inductive: that the formulation of scientific theory starts with the basic, raw evidence of the senses – simple, unbiased, unprejudiced observation. Out of these sensory data - commonly referred to as ‘facts’ – generalizations will form. The myth is that from a disorderly array of factual information an orderly, relevant theory will somehow emerge. However, the starting point of induction is an impossible one.

There is no such thing as unbiased observation. Regardless of how hypotheses arise, either guesswork or by inspiration, they can and must be tested rigorously, using the appropriate methodology. If the predictions you make as a result of deducing certain consequences from your hypothesis are not shown to be correct then you must discard or modify your hypothesis. If the predictions turn out to be correct then your hypothesis has been supported and may be retained until such time as some further test shows it not to be correct. This thesis concerns two types of research: **problem-solving**, and **exploratory**.

Problem-solving research starts from a particular problem, and it applies current knowledge to it, usually from different fields of expertise. The challenge in this type of research comes from rigorously defining the problem, and the method in which the solution will be sought. The use of Linked Data for sentiment and emotion analysis services and resources falls under this category. It is defined in terms of shortcomings of current approaches, and its hypotheses are formulated around real-world outcomes more so than on academic achievements.

On the other hand, we have exploratory research, where little is known about the problem, This type of research will need to examine what theories and concepts are appropriate,

¹To properly introduce the scientific methodology, we have borrowed heavily from Phillips, Pugh, et al., 2010, which we encourage any aspiring PhD student to read.

developing new ones if necessary, and whether existing methodologies can be used. It involves pushing the frontiers of knowledge in the hope that something useful will be discovered. The investigation of contextual information falls under this type of research, as it is a novel field of study, which combines several research areas. In this specific case, there is also a very precise end: improving sentiment and emotion analysis.

The following sections describe the phases in which we have organized our research, as well as the specific methodological considerations for each type of work (e.g., defining semantic vocabularies).

2.2 Phases

This thesis has been structured around three main fields of knowledge: semantic vocabularies, linked data services and social context. Each phase has been further divided into different tasks:

1. Phase 1 (**vocabularies**)

1. Study of semantic technologies and their application in research
2. Survey of currently used ontologies in the Sentiment and Emotion domain, as well as in the NLP community
3. Identification of missing definitions for Sentiment and Emotion Analysis
4. Definition of missing vocabularies
5. Publication of the vocabularies
6. Promoting the vocabularies and improving them based on community feedback

2. Phase 2 (**services**)

1. Study of the state of the art in semantic analysis services (e.g., NIF, Emotion-ML) and popular tools
2. Definition of a generic framework for sentiment and emotion analysis services
3. Development of a reference implementation
4. Adaptation of popular services, to serve as motivating example
5. Promotion of the tool and framework, and improvement based on feedback from the community

3. Phase 3 (**social context**)

1. Study of the state of the art in sentiment analysis with contextual information, and related social theories
2. Survey of works
3. Model of contextual information
4. Proposal of novel models for sentiment analysis with social context

2.3 Best practices for the publication of results

The expected results of this thesis include vocabularies, schemas, software architectures, reference implementations, surveys and classification models. All of these results need to be evaluated before being considered a scientific contribution (Section 2.4). Additionally, there are certain best practices that should be followed when publishing some of these types of results. In this section we will cover the publication of vocabularies and software.

These are the requirements for the definition of a vocabulary:

- Domain knowledge
- Thorough research to ensure no other vocabularies exist
- Identification of the core elements of the domain, and elements already modelled by other vocabularies
- Description of the domain using semantic technologies, in the form of a vocabulary
- Publication of the vocabulary and a set of examples to foster its use

There is no strong requirement to release vocabularies publicly, but doing so is highly recommended. It has several advantages, including more exposure, higher likelihood of re-utilization and modification based on feedback from the community. In consequence, the vocabularies in this thesis are public. When publishing a vocabulary we follow a set of rules: 1) the vocabulary should be published in at least one RDF format, preferably a human-readable format such as Notation3 or Turtle; 2) a documentation page with examples of use and the main concepts in the vocabulary is made, 3) the documentation and vocabulary are versioned, with a list of changes in each version, and a link to the previous version. Additionally, open vocabularies can be added to the Linked Open Vocabularies portal (Vandenbussche et al., 2017), a collection of curated reusable vocabularies for different domains. Adding your vocabulary to the LOV portal supports the project and makes it more likely for other researchers to re-use your vocabulary.

Regarding reference implementations and other forms of software results, we have taken measures to ensure that every piece of software is documented, is sufficiently tested, and behaves according to specification. To that effect, we have followed a Test-Driven Development approach, together with the Prototyping model. In practice, this means that features have been documented, tested, and added gradually to the implementation. We have used version control software (git) and an issue tracker. Each release of the software is tagged and automatically tested through a CI/CD (continuous integration, continuous delivery) pipeline. We have published all our software as Open Source on either GitHub or a public instance of GitLab. This enables a community-driven approach, where external researchers are encouraged to find bugs and send their modifications and improvements to the code. For documentation, we have mostly used auto-generated pages from the repository, hosted on ReadTheDocs (or similar). We have also distributed a packaged version of our libraries through the appropriate portals (PyPI for python). To ensure our software is compatible with as many environment as possible, and to simplify installation, we have also provided containerized images of our software. In particular, we have used Docker, the most popular open source containerization platform. Lastly, to promote the software in the research community, we have submitted a paper covering the software to an appropriate journal with an Original Software Track.

2.4 Evaluation

The expected results in this thesis fall under one of the following categories: vocabularies, software frameworks, reference implementations and classifiers (i.e., classification and predictive models).

Evaluating a vocabulary is very challenging. In our work, we will evaluate our vocabularies in two ways. First, we will evaluate the adequacy of each vocabulary by defining a series of scenarios that can not be represented using existing vocabularies, and trying to represent them using the vocabulary. The results of the representation will be tested for coverage, extensibility, and succinctness. This approach has an additional advantage: the scenarios used for evaluation also serve as documentation for the vocabulary. Secondly, our vocabularies are completely public and open to discussion, unlike other vocabularies in the literature. This allows us to receive public criticism, change requests and suggestions of new scenarios that are not properly covered by the vocabulary. We can then update the vocabularies based on that feedback, or open a discussion to cover the scenarios through other vocabularies.

Our evaluation of software frameworks is very similar to that of vocabularies. Namely,

we evaluate the coverage of a framework by its ability to cover different use cases. Its appropriateness and success is measured by its use in the community. To jumpstart the process, we also develop a reference implementation for each framework, which we apply in each use case. This enables us to assess the soundness of the framework, and the feasibility of its implementation. We also monitor for competing frameworks, and for the development of other implementations.

Reference implementations of a software framework are evaluated in different ways. First of all, a test suite is developed, following standard software engineering practices. Second of all, the implementation is tested in each of the use cases defined for the parent framework. In the case of plugin-based implementations, we also evaluate the ease of development and use of different types of plugins, as required by the use cases. The implementation is also released as open source in a public portal such as GitHub or GitLab, which encourages the community to find flaws and supply new use cases. Moreover, we require each reference implementation to be used and/or extended by at least three partners unrelated to the original development.

Lastly, classifiers are the easiest to properly evaluate, as evidenced by the literature. Each classifier is evaluated on multiple datasets, and several metrics, such as Accuracy, Precision and F-Score are calculated.