UNIVERSIDAD POLITÉCNICA DE MADRID

ESCUELA TÉCNICA SUPERIOR DE INGENIEROS DE
TELECOMUNICACIÓN

A DISTRIBUTIONAL SEMANTICS PERSPECTIVE OF
LEXICAL RESOURCES FOR AFFECT ANALYSIS: AN
APPLICATION TO EXTREMIST NARRATIVES

TESIS DOCTORAL

ÓSCAR ARAQUE IBORRA
Ingeniero de Telecomunicación

2020

UNIVERSIDAD POLITÉCNICA DE MADRID

ESCUELA TÉCNICA SUPERIOR DE INGENIEROS DE
TELECOMUNICACIÓN

A DISTRIBUTIONAL SEMANTICS PERSPECTIVE OF
LEXICAL RESOURCES FOR AFFECT ANALYSIS: AN
APPLICATION TO EXTREMIST NARRATIVES

TESIS DOCTORAL

ÓSCAR ARAQUE IBORRA
Ingeniero de Telecomunicación

2020

DEPARTAMENTO DE INGENIERÍA DE SISTEMAS
TELEMÁTICOS

ESCUELA TÉCNICA SUPERIOR DE INGENIEROS DE
TELECOMUNICACIÓN

UNIVERSIDAD POLITÉCNICA DE MADRID

A DISTRIBUTIONAL SEMANTICS PERSPECTIVE OF
LEXICAL RESOURCES FOR AFFECT ANALYSIS: AN
APPLICATION TO EXTREMIST NARRATIVES

AUTOR:

ÓSCAR ARAQUE IBORRA
Ingeniero de Telecomunicación

TUTOR:

CARLOS A. IGLESIAS
Doctor Ingeniero de Telecomunicación

2020

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 padre, a mi madre, y a todas las personas que me han ayudado en este camino.

Agradecimientos

Durante el desarrollo de la tesis me han acompañado un gran número de personas, y me gustaría darles a todas mi más sentido agradecimiento. Sin ellas no hubiera sido posible.

Quiero empezar con mi familia, que me ha apoyado desde incluso antes de empezar, dándome ánimos. A mis padres, que han fomentado mi carrera desde el inicio. A mis tíos y tías, que han estado para interesarse y darme fuerzas, y a mi abuela, que siempre tiene buenas palabras.

Claramente, a Cristina, que ha estado siempre a mi lado, ayudando en todo momento, y también porque me ha seguido en mis peripecias y viajes que, sin ella, no habrían sido iguales. Si hay una persona que ha tenido que aguantarme, es ella. Muchas gracias.

A Carlos Ángel, que ha sido un excelente tutor con el que he podido aprender y formarme. Esas largas charlas en el despacho han servido más de una vez y más de dos para meterle caña al trabajo, y también a mí mismo. También al resto del GSI: J, Álvaro, Sergio, por el día a día, las comidas y los cafés, sin olvidar el MK. A los que han pasado por los laboratorios y han hecho muchas tardes amenas: Manu, Alberto, Enrique y tantos otros. En Teleco, a todos los compañeros que me han acompañado en el viaje, siempre con una risa: Adri, Iván, Alfonso, y tantos otros.

No me quiero olvidar a la gente que me acogió en Milton Keynes, en especial a Miriam y a Martino, que hicieron de la estancia algo tremendamente agradable y divertido. Por último, le doy las gracias al DIT y a la ETSIT por haberme acogido y hacer del trabajo un placer.

No puedo olvidarme de la gente que me ha acompañado desde hace más de una década: Belén, Tomé, Rodri, David, Javi² y Floren. También quiero agradecerle el apoyo a toda la gente de televallecas: Lara, Álvaro, Raquel, Silvia, Sergio, Sandra, Kuba, Berta, Jose, y todos los demás.

A todos, mil gracias.

Resumen

Los campos de Análisis de Sentimientos y Emociones son prominentes en el Procesado de Lenguaje Natural (NLP, en inglés). Modelar emociones requiere de un número de técnicas que, en la mayoría de los casos, se comparten con otras áreas del NLP. Es este contexto, esta tesis aborda el desarrollo de métodos de aprendizaje automático novedosos a través de la combinación de características superficiales y profundas en Análisis de Sentimientos y Emociones. Tras observar los resultados obtenidos, los métodos desarrollados han sido adaptados a otras áreas del NLP, transfiriendo así los resultados obtenidos a campos que se pueden beneficiar en gran manera de dichas técnicas.

De esta manera, hemos diseñado una taxonomía que permite clasificar diferentes tipos de técnicas en Análisis de Sentimientos, según las características y combinaciones usadas. Esta taxonomía nos ha permitido desarrollar varios modelos de aprendizaje automático que combinan características superficiales y profundas, además de un variedad de modelos de aprendizaje, sobre todo las técnicas de aprendizaje profundo. Estos modelos han sido evaluados mediante en entornos de Análisis de Sentimientos, a niveles de documento y aspecto. A continuación, hemos contribuido con una técnica para la generación de léxicos de sentimientos específicos del dominio, usando el algoritmo de retro-propagación. En una línea similar de trabajo, también hemos extendido un método para la generación de léxicos, obteniendo un recurso bilingüe anotado en Inglés e Italiano.

La contribución central de esta tesis es un modelo que explota la similitud semántica en representaciones distribuidas a través de léxicos, extrayendo características mediante el uso de *word embeddings* y un léxico de sentimientos: el modelo SIMON. Hemos observado que SIMON ha mostrado resultado positivos en la evaluación experimental, y ofrece una alternativa potente a las técnicas clásicas de uso de léxicos.

A la luz de esta última contribución, esta tesis estudia la adaptación de los métodos desarrollados a otros campos. Más concretamente, esta tesis ha contribuido a la (a) detección de radicalización, donde hemos adaptado el modelo SIMON, combinándolo con un método de extracción de características emotivas; y a la (b) estimación de valores morales, en la que se ha usado SIMON para emplear un nuevo léxico que hemos generado.

Abstract

Sentiment and Emotion Analysis are prominent fields in Natural Language Processing (NLP) and have contributed to its progress. Modeling emotions requires a number of techniques and methods that, in most of the cases, are shared with other NLP areas. In such a context, this thesis addresses the development of novel machine learning methods through the combination of both surface and deep features for Sentiment and Emotion Analysis. Observing the results obtained, the developed methods have been adapted to other NLP areas, transferring the obtained results to fields that can largely benefit from such novel techniques.

In this way, we have designed a taxonomy that classifies different approaches in Sentiment Analysis, attending to the features and combinations used. This taxonomy has allowed us to develop several machine learning models that combine both surface and deep features, as well as a variety of learning models, with an especial focus on deep learning approaches. These models have been evaluated in document and aspect-level Sentiment Analysis frameworks. Following, we contributed with a novel technique for generating domain-specific sentiment lexicons through the backpropagation algorithm. In a similar line of work, this thesis also extends a method for generating emotion lexicons, obtaining a bilingual resource with emotion annotations in both English and Italian.

The thesis' core contribution is a similarity-based perspective on lexicons that extracts features exploiting both a word embedding model and a sentiment lexicon: the SIMilarity-based sentiment projectiON (SIMON) model. We have observed that SIMON has shown positive results in the experimental evaluation, offering a compelling alternative to classical lexicon usage techniques.

In light of this last contribution, this thesis studies the adaptability of methods that are developed in the context of Sentiment and Emotion Analysis to other fields. More concretely, this thesis has contributed to radicalization detection and moral value estimation. In the area of radicalization detection, we have adapted the SIMON model, combining it with an emotion-driven feature extraction method. Similarly, for moral value estimation, SIMON has been used to exploit a novel lexicon we have generated.

Contents

Resumen	I
Abstract	III
Contents	V
Reader's guide	1
1 Introduction	3
1.1 Motivation	4
1.2 Background	10
1.2.1 Lexicons	10
1.2.2 Word Embedding Models	15
1.2.3 Semantic Similarity	18
1.2.4 Affect Analysis of Online Radicalism	19
1.2.5 Moral Foundations Theory	24
1.3 Hypotheses	26
1.4 Objectives	27
1.5 Document outline	28
2 Methodology	29
2.1 Introduction	30
2.2 Machine Learning Experimentation	31
2.3 Evaluation	33

2.4	Statistical Analysis	34
3	Publications	37
3.1	Introduction	38
3.2	List of central publications	38
3.2.1	An Approach for Radicalization Detection Based on Emotion Signals and Semantic Similarity	38
3.2.2	MoralStrength: Exploiting a moral lexicon and embedding similarity for moral foundations prediction	55
3.2.3	A Semantic Similarity-based Perspective of Affect Lexicons for Senti- ment Analysis	67
3.2.4	DepecheMood++: a Bilingual Emotion Lexicon Built Through Simple Yet Powerful Techniques	82
3.2.5	Enhancing Deep Learning Sentiment Analysis with Ensemble Tech- niques in Social Applications	95
4	General Discussion, Conclusions and Future Research	107
4.1	Overview	108
4.2	Scientific results	110
4.2.1	Objective 1: Surface and Deep methods framework	110
4.2.2	Objective 2: Definition of a lexicon generation methodology	114
4.2.3	Objective 3: Definition of a similarity-based model over distributed representations	119
4.2.4	Objective 4: Transference of a similarity-based model that exploits lexicons to other domains	125
4.2.5	Other	130
4.3	Software Publication	131
4.4	Conclusions	132
4.5	Future Research	137

Bibliography	141
List of Figures	159
List of Tables	161
Glossary	163
Appendix A Publications	165
A.1 Summary of all publications	165
A.2 Rest of publications	167
A.2.1 Radical Text Detection based on Stylometry	167
A.2.2 Senpy: A Framework for Semantic Sentiment and Emotion Analysis Services	176
A.2.3 GSI-UPM at SemEval-2019 Task 5: Semantic Similarity and Word Embeddings for Multilingual Detection of Hate Speech Against Immi- grants and Women on Twitter	180
A.2.4 How Well Do Spaniards Sleep? Analysis of Sleep Disorders Based on Twitter Mining	189
A.2.5 Neural Domain Adaptation of Sentiment Lexicons	198
A.2.6 Applying Recurrent Neural Networks to Sentiment Analysis of Spanish Tweets	205
A.2.7 Mining the Opinionated Web: Classification and Detection of Aspect Contexts for Aspect Based Sentiment Analysis	212
A.2.8 Senpy: A Pragmatic Linked Sentiment Analysis Framework	221
A.2.9 Aspect Based Sentiment Analysis of Spanish Tweets	230

Reader's guide

This thesis is presented in the form of a compendium of publications. That is, this document does not pretend to be a full monograph. Instead, the essence of this document lies in the collection of publications by the author. Such publications must comply with the following requirements: the Ph.D. student must appear as the first author, and at least three of these publications must be published in Q1 or Q2 ranked journals, as indexed by JCR or Scopus. These conditions are satisfied, as can be observed in Chapter 3 and Appendix A. Following the guidelines by Universidad Politécnica de Madrid, the structure of this document must comply with a specific structure. Thus, this section intends to guide the reader into this document.

Chapter 1 introduces the main topics to which this thesis revolves. In this way, it details the motivation (Section 1.1) and background (Section 1.2) that have driven the research done during the Ph.D. Following, this chapter also discovers the hypotheses (Section 1.3) and objectives (Section 1.4) that we have followed. These objectives and hypotheses have guided our activities.

In Chapter 2, the full methodology of the thesis is detailed. This chapter describes how we have performed experimentation and model evaluations. It also describes the best practices for machine learning experimentation (Section 2.2), the evaluation methods (Section 2.3), as well as the use of statistical analysis (Section 2.4), a field that has been central to this thesis. These techniques define how we have operated during the accomplished research and how posterior conclusions can be taken from the obtained results.

Chapter 3 contains the central publications of this thesis. Each paper is presented in a separate section to facilitate reading, and it includes a preliminary table with relevant details (e.g., authors, title, journal, impact factor). Throughout the document, there are several publication tables, and all of them refer to the section where the paper can be read in full. Please note that in Chapter 3, not all publications by the Ph.D. student are included. This is done following the guidelines by Universidad Politécnica de Madrid, as this chapter must only include the publications that fulfill the described requirements. A full collection of publications can be found in Appendix A.

Chapter 4 contains a general discussion about the thesis. Firstly, we present an overview of the results obtained during the development of the thesis (Section 4.1). Once those have been described, the scientific results are outlined in relation to the previously presented thesis' objectives (Section 4.2). This chapter also describes the software tools that have emerged as part of the research process (Section 4.3). Finally, the chapter concludes with the main conclusions gathered (Section 4.4), and with some work lines that can be addressed in the future (Section 4.5).

As mentioned, Appendix A contains all the publications of this thesis. A complete table can be found in Section A.1, while Section A.2 contains the full texts of the publications that do not fulfill the university requirements. We would like to stress that several publications have contributed significantly to this thesis and are discussed throughout the document, although do not appear in Chapter 3.

CHAPTER 1

Introduction

Research requires much effort in preparation and documentation. Knowing previous research is vital and must not be ignored. This chapter discusses the context of this Ph.D. thesis, which includes its motivation, relevant background, objectives, and hypotheses.

“On the Shoulders of Giants.” A popular search engine.

1.1 Motivation

The consistent growth of users and user-generated content in many web sites, social networks, and online consumer platforms such as Twitter, Amazon, and Yelp has incremented the quantity of opinionated information available on the Internet. This tendency, combined with the fast-spreading nature of content online, has turned online user-generated content into a precious asset. Said information has inherent value for many online and offline businesses, as well as research entities, since language usage reflects our opinions, thoughts, emotions, values, and culture, as we communicate with others. Online opinions can, for example, directly affect future sales in a wide range of fields such as hotels, e-commerce, restaurants, and electronics (Chevalier et al., 2006; F. Zhu et al., 2010). Consequently, interest in Natural Language Processing (NLP) techniques as a way of automatically extracting and analyzing user-generated content has risen (Liu, 2015).

In this context, many NLP methods are being developed and applied in order to analyze the massive quantity of information that is available on the Internet. In particular, this thesis revolves around Sentiment Analysis (SA) and Emotion Analysis: two increasingly growing tasks (Liu, 2015) whose goals are the study of opinions, sentiments, and emotions expressed in text generated by a human party. In Sentiment Analysis works, text is generally annotated into several categories, being *positive*, *negative*, and *neutral* the most common. These annotations can be associated with a numerical sentiment score, usually in the $[-1, 1]$ range.

Since the study of opinions in text comprises a broad range of possibilities, Sentiment Analysis tasks can be classified into three categories, attending to the granularity level (Liu, 2015): *document level*, as in a movie review that can span one or more paragraphs; *sentence level*, where sentiment is bounded to a single sentence of text; and *aspect level*, in which case sentiment is associated with a word or small group of words (e.g., *good food*, *short battery duration*).

In comparison, Emotion Analysis represents an extension over the negative/positive distinction that is done in Sentiment Analysis, considering individual emotions rather than a binary label. In this way, among the most common emotions that are measured, we can find *anger*, *disgust*, *joy*, *fear*, *sadness* and *surprise*, which roughly correspond to the ones described in the work of Paul Ekman (Ekman, 1992). In such a light, one can consider Emotion Analysis as an extension over Sentiment Analysis (Staiano et al., 2014), both belonging to Affect Computing.

Even though both Sentiment Analysis and Emotion Analysis represent different research

problems, both share many challenges, methodologies, and, in general, approaches. As such, and in order to facilitate the reading, many of the descriptions that appear in the rest of the section apply both to Sentiment and Emotion Analysis, even if only sentiment is mentioned.

Due to the subjective nature of Sentiment Analysis and Emotion Analysis, it is of no surprise that sentiment and emotion words are a key indicator of subjective polarity. For instance, *good*, *great* and *brilliant* can be considered positive words while *bad*, *worse* and *disastrous* can express negative attitudes. Consequently, sentiment lexicons, which gather sentiment words, are used extensively by the research community (Ravi et al., 2015a). Sentiment lexicons can be organized in three types, attending to which information is contained in them (Wu et al., 2013; Peng et al., 2018): (i) those who contain only sentiment words, a list of sentiment-bearing words; (ii) the ones that are formed by both sentiment words and polarity orientations, a list of words with only positive and negative annotations; and (iii) the lexicons that offer sentiment words with orientation and intensity (Taboada et al., 2011), a list of words with scalar numerical values.

The use of lexicon-based techniques has a number of advantages (Taboada et al., 2011). First, the linguistic content can be taken into account through mechanisms such as sentiment valence shifting (Shanahan et al., 2006) considering both intensifiers (e.g., *very bad*) and negations (e.g., *not happy*). In addition, sentiment orientation of lexical entities can be differentiated based on their characteristics. Moreover, language-dependent characteristics can be included in these approaches. Nevertheless, lexicon-based approaches have several drawbacks: the need of a lexicon that is consistent and reliable (Taboada et al., 2011), as well as the variability of opinion words across domains (Peter D Turney, 2002), contexts (Ding et al., 2008), and languages (Perez-Rosas et al., 2012). These dependencies make it hard to maintain independent domain lexicons (Qiu et al., 2009).

The most common approach that makes use of sentiment or emotion lexicons is keyword matching, also called keyword spotting (Cambria, Schuller, et al., 2013). Basically, this technique involves detecting the presence of certain sentiment bearer words, thus obtaining the sentiment estimation as an aggregate of the associated sentiment values. Although this method is certainly simple and computationally cheap, it is also limited, as it happens in the case of domain adaptation. Sentiment words can display variations in their polarity values depending on the contextual domain (Peter D Turney, 2002), language (Perez-Rosas et al., 2012), or even context (Ding et al., 2008), causing lexicon-based approaches to decrease their classification performance. An additional challenge that arises in relation to the use of sentiment lexicons is the selection between the variety of these resources. It is still an open question how to select and process sentiment and emotion lexicons (Bravo-Marquez

et al., 2014), as well as how this selection affects the performance of sentiment and emotion classification tasks.

As discussed, exploiting sentiment and emotions lexicons constitutes a research problem that is being actively studied. Similarly, obtaining high-quality lexicons is an active subject of research (S. M. Mohammad and Peter D. Turney, 2010). Generally speaking, lexicon acquisition has been implemented through three distinct methodologies: by manual creation (e.g., crowdsourcing annotations), by a dictionary-driven method, or by automatic derivation from already annotated corpora. While the two first approaches provide more precise lexicons, the latter usually grants a higher coverage. Regardless of the chosen approach, when used as baselines or as additional features for more complex learning models, lexicons’ quality is not usually evaluated. That is, the performances used for comparison are rather weak, a fact that could be arguably seen to slow progress in the field.

Additional challenges that arise when trying to obtain sentiment or emotion lexicons are related to the lexicons’ annotations. It is worth taking into account the fact that, as discussed previously, words can vary their sentiment or emotion values across different uses of the language, time, and domain (Hamilton et al., 2016). Also, even the selection of sentiment or emotion words are dependent on the domain (Peter D Turney, 2002). For these reasons, the task of automatically generating domain-independent and domain-adapted lexicons flexibly represents a significant research problem.

While sentiment and emotions lexicons are indeed a relevant resource frequently used in the field, nowadays the predominant approaches rely on *machine learning* techniques (Pang et al., 2002; Read, 2005; Wang et al., 2012). Traditional approaches use the Bag Of Words (BOW) method, where a document is mapped to a numerical feature vector, which offers a machine learning algorithm a unified representation of the analyzed text. Although the Bag Of Words approach is quite popular, simple, and efficient, a fraction of the information from the original natural language is lost (Xia et al., 2010); e.g., word order is disrupted, and syntactic structures are broken. Accordingly, to improve the performance offered by the BOW method, various types of features have been exploited, such as higher-order *n-grams* (Pak et al., 2010). Another type of feature that can be used is Part-of-Speech (PoS) tagging, which extracts information along the lines of syntactic analysis, as described in Gimpel et al. (2010). Some authors refer to this kind of features as surface forms, as they consist of lexical and syntactical information that relies on the patter of the text, rather than on its semantic aspects. In this thesis, we follow such a nomenclature.

In practice, extracting complex features from text, examining which features are relevant, and selecting a learning algorithm are fundamental questions that are present in the

machine learning methods (A. Agarwal et al., 2011; Sharma et al., 2012; Wilson et al., 2009). Mentioned traditional approaches rely on manual feature engineering, which is costly and time-consuming.

As advancement in the effort to effectively obtain text representations, we find *distributed representations*, or more concretely word embeddings (Mikolov et al., 2013; Le et al., 2014), which are considered a relevant characteristic of state-of-the-art sentiment and emotion analysis systems.

These techniques encode text into fixed-length vectors that can be used directly by machine learning methods, being neural networks the most used recently (Collobert et al., 2011). The use of (deep) neural networks enables the system to learn complex features automatically extracted from data, minimizing manual domain-oriented efforts Bengio, 2009; Alpaydin, 2014. Lately, neural networks have shown excellent performance in Natural Language Processing tasks, including sentiment and emotion analysis (Collobert et al., 2011).

The main idea of neural techniques is to learn complex features extracted from data with minimum external contribution (Bengio, 2009) using deep neural networks (Alpaydin, 2014). Contrary to the mentioned traditional methods, neural learners do not need to be passed manually crafted features: they automatically learn new complex features.

Nevertheless, a requirement of these approaches is the use of large amounts of data, normally annotated (Mikolov et al., 2013). Both automatic feature extraction and availability of resources are very important when comparing the traditional machine learning approach and novel deep learning techniques. Efficiently training neural networks in small datasets is still an open challenge, whose resolutions could improve sentiment and emotion analysis systems.

However, it is not clear whether the specialization capacity of traditional approaches towards a domain can be surpassed with the generalization capacity of deep learning-based models in all NLP tasks, or if it is possible to successfully combine these two techniques in a wide range of applications. Besides, when attending to the application of neural methods to sentiment and emotion analysis, subjective prior knowledge is still a key asset. This takes us back to the effective use of sentiment and emotion lexicons along with neural architectures.

While some core contributions of this thesis are framed in sentiment and emotions analysis, during the development of the thesis a number of advancements have been made in other fields. In this way, the foundations, techniques, and models that have been developed and studied can be applied to solve other problems that fall outside of the sentiment and emotion estimation of texts. An overview of the domains where such techniques have

been applied can be seen in Figure 1.1. In particular, this thesis has contributed to the study and automatic analysis of texts concerning radical narratives, and Moral Foundations Theory (MFT).

Application of Natural Language Processing techniques to social science, psychology, and other fields is common and can yield exciting results (Crowston et al., 2012; Lin et al., 2018; Fernandez et al., 2018). It is clear that the growing amount of human-generated text is of particular interest to social scientists. Still, interesting as this data for scientists may be, it is also clear that single researchers can not manually analyze even a portion of all the available materials. Sampling, analyzing, and categorizing requires a fair amount of domain knowledge. Thus, technologies and methods that support researchers with these massive problems are becoming more important and are nowadays widely used (Wiedemann, 2013).

In this line of thought, as mentioned, this thesis has contributed to different fields by applying, adapting, and designing methods that were initially developed for affect analysis. Also, in some cases, fundamental sentiment or emotion ideas have been adapted to contribute to such fields.

As motivation, we would like to emphasize the application to radicalization detection firstly. Nowadays, modern terrorist organizations are adapting and changing in order to achieve their objectives, overcoming counter-terrorist actions. In this sense, terrorists are effectively using the Internet as part of their propaganda and recruitment strategies (Birmingham et al., 2009). Consequently, online radicalization detection represents a significant concern for counter-terrorists organizations, Law Enforcement Agencies, and the populations in general. Digital traces left on social networks can be used as social signs to detect online radicalism, based on text mining and social analysis techniques (Birmingham et al., 2009), which analyze different aspects of these interactions such as sentiment, emotions, topics, and the social context of users and semantics (Saif et al., 2017).

Additionally, social science studies have highlighted the role of emotions for understanding radicalization (Stekelenburg, 2017; Rice, 2009). A previous work (Stekelenburg, 2017) argues that the radicalization process is fueled by the emotions of anger, contempt, and disgust. Considering these observations, we have detected an interesting application that combines emotion information with domain-aware text representations for radicalization detection in social networks.

Although radicalization is indeed an exciting application field, an additional field where computational approaches can be utilized in social sciences is the study of morality. Moral values influence the way we rationalize and form an opinion upon controversial topics (Sagi et al., 2014b). They are also closely related to our political views (Miles et al., 2015) that

may affect immigration (Grover et al., 2019), political extremism (Alizadeh et al., 2019), and partner violence (Vecina et al., 2015). Also, it has been previously studied how moral values can be used to predict violent protests based on user-generated text (Mooijman et al., 2018). Moral Foundations Theory represents the most established theory in psychology and social sciences. It is also being adopted in the computational social science field since it defines a clear taxonomy of moral values by means of the Moral Foundations Dictionary (MFD) (Graham et al., 2009), which is an essential resource for Natural Language Processing applications.

Concerning the Moral Foundations Dictionary resource, Graham et al. (2009) highlight the difficulty of creating such a resource since linguistic, cultural, and historical context reflect on language usage. Considering the most important limitations of the Moral Foundations Dictionary, we find: (i) a limited amount of terms (lemmas and stem of words); (ii) terms that are rarely used in everyday language (e.g., *homologous*, *apostasy*); (iii) a binary annotation to a bipolar scale, called vice and virtue, that does not indicate strength. In this context, a resource that offers more coverage and an improved numerical annotation could significantly foster research in this area.

In summary, although sentiment and emotion analysis benefits from advancements in Natural Language Processing techniques like distributed representations, many challenges remain in relation to the combination of subjective knowledge with advanced neural networks models.

Models that efficiently exploit these resources while maintaining a low complexity will surely benefit the field. Adaptation of such methods to specific domains is also a challenge, including its generation. Additionally, the use of methods and models initially developed for sentiment and emotion analysis is an exciting field that can benefit both affective computing and social sciences. Adapting and improving these models to function under such conditions opens new research problems.

Figure 1.1: Overview of the research fields where this thesis has made contributions. Core contributions are delineated stronger.

1.2 Background

1.2.1 Lexicons

Previously, we have described the importance of lexicons in Sentiment Analysis. How to effectively use a sentiment lexicon is a common problem that appears in the vast majority of Sentiment Analysis works. An interesting survey that tackles the use of sentiment lexicons for Sentiment Analysis is done in (Ravi et al., 2015b). Numerous sentiment lexicons can be used, and determining which characteristics determine the final performance of a system that uses them is an open challenge. In this sense, the work in (Bravo-Marquez et al., 2014) treats some of these issues, analyzing several sentiment lexicons, and how they can be complemented in a Sentiment Analysis system.

One relevant attribute of sentiment lexicons is the creation methodology, which can be divided into three main categories (Liu and L. Zhang, 2012; Araque, Guerini, et al., 2017): (i) manual approach, (ii) dictionary-based approach, and (iii) corpus-based approach. The manual approach is made by humans, resulting in a very time-consuming process; thus, lexicons generated in such a manner are typically combined with the other two, that make use of automatic strategies. Many dictionary-based methods involve bootstrapping from a small set of opinionated words, using them as seeds to search related words in a known dictionary such as WordNet (Miller, 1995) or SentiWordNet (Baccianella et al., 2010). Traditionally, dictionary-based approaches are not able to obtain domain-specific orientations (Liu and L. Zhang, 2012). Although, in order to generate domain-adapted dictionaries, some recent

works use unsupervised pattern-based approach (Agathangelou et al., 2018), as well as morphosyntactic information (Vulcu et al., 2014). Despite this, corpus-based techniques can be used to alleviate the issue mentioned above. These last type of lexicon generation methods rely on co-occurrence patterns detected on an unsupervised corpus, and a seed set of opinion words to locate opinionated words in the corpus (Yadollahi et al., 2017; Hamilton et al., 2016).

As done in this thesis, sentiment lexicons have been used as features in supervised machine learning scenarios. In (Kouloumpis et al., 2011), lexical resources are complemented with n-gram, Part-of-Speech (PoS), and micro-blogging specific features, such as the presence of emoticons. In this way, the authors show that both lexicon and micro-blogging features are useful in their validation. Similarly, the work in (S. Mohammad, Kiritchenko, and X. Zhu, 2013) makes use of two sentiment lexicons, combining this information with n-grams, PoS. All these features are used by an SVM-based classifier, resulting in a state-of-the-art system in a Sentiment Analysis competition. In some cases, lexicon features have been complemented with domain-specific features, as in (Jiang et al., 2011), where the search query determines the domain. Also, several lexicons can be used at the same time, as in the case of (Zirn et al., 2011). This last work integrates several lexicons using Markov logic with information about relations between neighboring words.

While many sentiment lexica have been produced, fewer linguistic resources for emotion research are described in the literature. One of the first is WordNet-Affect (Strapparava and Valitutti, 2004), a manually-built extension of WordNet. About 1k lemmas are assigned a label taken from a hierarchy of 311 affective labels, including Ekman’s six emotions (Ekman and Friesen, 1971). AffectNet (Cambria, Mazzocco, et al., 2011) is a semantic network containing about 10k items; emotions are represented in a vector space where the dimensions are “pleasantness”, “attention”, “sensitivity”, and “aptitude”. Each of these dimensions can have a value between -3 and +3 for a total of 24 distinct emotions that can be combined.

The Affect database (Neviarouskaya et al., 2010) contains 2.5k lemmas taken from WordNet-Affect and has been manually enriched by adding the strength of association with Izard’s basic emotions (Izard, 1977). EmoLex (S. Mohammad and P. Turney, 2013) is a crowdsourced lexicon containing 14k lemmas, each annotated with binary associations to Plutchik’s eight emotions (Plutchik, 1980).

A recent resource is the NRC Affect Intensity Lexicon (S. Mohammad, 2018), which includes 6k entries manually annotated with joy, fear, anger, and sadness. In a similar line of work, the NRC Hashtag Emotion Lexicon (S. Mohammad and Kiritchenko, 2015) has a coverage of over 16k unigrams. This resource has been automatically inferred

from microblogging messages distantly annotated by emotional hashtags. As such, this lexicon is particularly useful when applied to the Twitter domain.

In (Bandhakavi et al., 2017), a generative unigram mixture model is proposed as a solution for creating domain-specific emotion lexica by jointly modeling “emotionality and neutrality of words”, starting from a corpus of labeled documents.

Apart from theories of basic emotions, the literature has also proposed several non-categorical models, where emotions are represented as points in a vector space. The Valence-Arousal-Dominance (VAD) model, in particular, is worth mentioning since multiple resources are based on it. ANEW (Bradley et al., 1999) is a set of 1k words manually annotated. A crowdsourced extension of it (Warriner et al., 2013) provides fine-grained VAD ratings for 14k frequent-usage words. A recent work, the NRC Valence, Arousal, Dominance Lexicon (S. M. Mohammad, 2018), contains 20k terms manually annotated using Best-Worst Scaling (Kiritchenko et al., 2016). While many of these resources were used for Sentiment Analysis, considering only the valence scores (which roughly corresponds to the “sentiment” or “polarity” of a word), the VAD model actually represents emotions.

A recent work has demonstrated the possibility of seamlessly converting VAD annotations between Ekman’s categorical emotions, and vice versa (Buechel et al., 2018). Recently, a multilingual study using the GRID dictionary extended the VAD model to include a new dimension, Novelty, but no emotion lexicon is using this model yet (Fontaine et al., 2013).

As part of this thesis, the DepecheMood lexicon is extended **“MoralStrength: Exploiting a moral lexicon and embedding similarity for moral foundations prediction”** (Section 3.2.2). The original DepecheMood lexicon (Staiano et al., 2014) and its extension, are the largest among these emotion lexica. They are automatically derived from a corpus of news articles with annotated emotions, and thus they can be assumed to be general-purpose. Their set of emotions is also peculiar: while most of the other lexica are related or easily mapped to Ekman’s, Plutchik’s or Izard’s sets, the categories “annoyed” and “amused” are unique for these resources.

As a summary, Table 1.1 shows a comparison of these emotion lexica. The “Annotation type” column indicates whether a lexicon reports binary values for an emotion, or a number indicating its strength. Similarly, the “Annotation process” column indicates whether the generation process is manual or automatic (e.g., obtained as a result of an algorithm).

In a parallel effort in the field, many works have been oriented to domain adaptation of sentiment lexicons. Given the idiosyncrasy of sentiment traits in a given domain, a necessity for prior domain knowledge arises. In this thesis, we understand *domain-oriented*

Domain	Name	No. entries	Annotation type	Annotation process	Reference
Sentiment	General Inquirer	4k	Categorical	Manual	Stone et al. (1966)
	ANEW	1k	Numerical	Manual	Bradley et al. (1999)
	-	6k	Categorical	Automatic	Hu et al. (2004)
	SentiWordNet	117k	Numerical	Automatic	Baccianella et al. (2010)
	AFINN	2k	Numerical	Manual	Nielsen (2011)
	SO-CAL	4k	Numerical	Manual	Taboada et al. (2011)
	-	14k	Numerical	Manual	Warriner et al. (2013)
	SentiWords	144k	Numerical	Automatic	Gatti et al. (2016)
Emotion	Fuzzy Affect	4k	Numerical	Manual	Subasic et al. (2001)
	NRC VAD	20k	Numerical	Manual	S. M. Mohammad (2018)
	WordNet-Affect	1k	Categorical	Manual	Strapparava and Valitutti (2004)
	Affect database	2k	Numerical	Manual	Neviarouskaya et al. (2010)
	AffectNet	10k	Categorical	Automatic	Cambria, Mazzocco, et al. (2011)
	EmoLex	14k	Categorical	Manual	S. Mohammad and P. Turney (2013)
	NRC Hashtag	16k	Numerical	Automatic	S. Mohammad and Kiritchenko (2015)
	NRC Affect	6k	Numerical	Manual	S. Mohammad (2018)

Table 1.1: Overview of related Sentiment and Emotion lexica.

or *domain-adapted* lexicons as resources that have, in one way or another, specific domain knowledge.

There is a great variety of statistical information extraction methods for the generation of domain-oriented lexicons. Nevertheless, some works share some common methodology, as is the case of *optimization-based* approaches. For instance, the work shown in Du et al. (2010) proposes an information bottleneck problem that blends cross and within-domain knowledge. This problem is then solved with an iterative optimization method that, upon convergence, obtains the desired lexicon. Choi et al. (2009) explores a different approach, where a linear programming technique is aimed to directly reflect the characteristics of a particular domain. This method exploits the relation between words and opinion expressions to compute the most likely polarity value in that domain. Another explored technique is presented in Lu et

al. (2011), in which several opinion signals are extracted from an unlabeled opinionated text collection. These signals are combined into an optimization problem and then transformed into a linear programming formulation by means of a mathematical transformation. In this way, the method can learn the sentiment of words and aspects.

Another common technique used when generating domain-specific lexicons is *label propagation*. In this line of work, the technique explained in Qiu et al. (2009) introduces a method called double propagation. This method consists of several extraction rules that are designed based on dependency trees and exploit the relations between modifier sentiment words and their associated topics, as well as the sentiment and topic words. Similar research (Huang et al., 2014) proposes an automatic construction strategy based on constrained label propagation. The strategy firstly extracts candidate sentiment terms, and then it uses domain and morphological constraints to spread to the entire set of candidates, improving the lexicon. Label propagation can also be applied to different structures, such as a lexical graph. This approach is taken in Hamilton et al. (2016), where a set of high-quality word embeddings are transformed into a lexical graph by connecting each word with its nearest neighbors in the semantic space. Then, the sentiment labels are propagated from a set of seed words through the graph using a random walk algorithm.

Lastly, other methods directly extract statistical information from corpora and transform this knowledge into a domain-adapted sentiment lexicon. The method described in Almatarneh et al. (2017) uses frequency information to assign different weights to positive and negative words in the domain of movie reviews. Another example of such approaches is Demiroz et al. (2012), which uses the *TF-IDF* score for each sentiment polarity to adapt an existing general lexicon to a specific domain. Also, the works described in Staiano et al. (2014) and Song et al. (2016) describe the *Compositional Semantics* methods that, parting from a frequency measure of the words in the corpus and a set of emotion labels (this could be used in sentiment as well), builds a lexicon from a word-emotion matrix.

This thesis addresses lexicons in various ways. We have seen how extended utilization methods are straightforward, but lack in representational power. Following this research line, this thesis has contributed with the work described in **“A semantic similarity-based perspective of affect lexicons for sentiment analysis”** (Section 3.2.3). Another challenge lies in lexicon generation. Automatically deriving a lexicon that has both high coverage and quality annotations is a difficult task, and many works address this issue. In this line, this thesis offers two approaches, one that exploits a known neural network algorithm (**“Neural domain adaptation of sentiment lexicons”** (Section A.2.5)), and another that uses a simple and effective statistical method (**“DepecheMood++: a Bilingual Emotion**

Lexicon Built Through Simple Yet Powerful Techniques” (Section 3.2.4)).

1.2.2 Word Embedding Models

Continuous word vector representations contain syntactic and semantic regularities present in the language, expressed as relation offsets in the vector space (Mikolov et al., 2013). Most common approaches train unsupervised word embeddings models without a specific objective, but rather to capture language knowledge. This type of word vectors, which are trained using co-occurrence information, is typically called generic or pre-trained word vectors. In this thesis, pre-trained word vectors are used for all types of proposed feature extraction methods.

In the same way that bag-of-words features are exploited for textual representation in Sentiment Analysis, word embeddings can be similarly used (Mikolov et al., 2013). Some straightforward approaches that use embeddings as features for classification have been studied. The work contained in (Schnabel et al., 2015) studies the effectiveness of using word embeddings applied to several tasks, one of them being Sentiment Analysis. This study offers an overview of unsupervised embedding techniques and how they can obtain meaningful text representations. In the work by D. Zhang et al. (2015), an SVM classifier is trained over embedding representations to predict sentiment polarity. Through their results, the authors argue that embeddings contain deep semantic features between words, which can be positive for Sentiment Analysis. On top of this, word embeddings based techniques are extensively used in public challenges where competitors aim at obtaining the highest scores in the task of Sentiment Analysis (Nakov et al., 2016; Rosenthal et al., 2017). An interesting work that uses embedding with special consideration to the reduction the computational complexity is shown in (Joulin et al., 2016); curiously enough, the authors report a similar accuracy to that of more complex neural models in several tasks, including Sentiment Analysis. For a more detailed description of how to train and use word embeddings, we refer to (Lai et al., 2016).

One relevant limitation of the presented word embedding models is that they ignore the morphology of words, assigning a distinct vector to each word. In light of this, newer models have been proposed that exploit additional sources of information in order to obtain more robust word representations. Bojanowski et al. (2017) propose the *FastText* word embedding model, where each word is represented by a character n -gram. In this way, a vector representation for a specific word is the result of the sum of the vectors that represent the character n -grams that form the said word. This method trains fast and has the advantage that allows computing representations of words that have not been previously

seen in the training data. In a complementary work (Joulin et al., 2016), the authors show how this kind of embedding model can reach high performances in several NLP tasks, including Sentiment Analysis.

A similar work is presented in Peters et al. (2018), introducing the Embeddings from Language Models (ELMo) model. This model uses both complex characteristics of word use (e.g., semantics and syntax) and variations across linguistic contexts. These word representations are learned from the internal states of a specialized neural network, pre-trained on a large text corpus. Such embeddings can be easily added to a learning system, and that they can boost the performance on significant NLP problems, being one of those Sentiment Analysis.

In the line of using embedding-based approaches, many authors have proposed complex neural architectures that are able to exploit word vectors successfully. One architecture family is Recurrent Neural Networks, which can adapt to different sequence sizes; these results especially useful in text processing, as a document can be modeled as a sequence of sentences, words, or even characters. As shown by Socher et al. (2013), the study of compositionality in sentiment classification tasks has proven to be relevant. This work proposes the Recursive Neural Tensor Network (RNTN) model, and it also shows how RNTN outperforms previous models on both fine-grained and binary Sentiment Analysis tasks.

The RNTN model represents a phrase using word vectors composed from the structure given by a parse tree, computing vectors for higher nodes in the tree using a tensor-based composition function. Another interesting work (Dai et al., 2015) tackles the use of Long Short-Term Memory (LSTM) networks, which read the input sequence into a vector. These vectors are used then to predict the sentiment label. A similar approach (Tai et al., 2015) leverages the use of both LSTM and parse trees, outperforming previous systems and strong baselines in sentence relatedness and sentiment classification tasks. As an additional technique, attention mechanisms are normally used in recurrent architectures (Bahdanau et al., 2014), which allows models to search for parts of the input relevant to the problem. Building on top of this, the work in Yang et al. (2016) incorporates knowledge of document structure, developing an attention-based approach under the assumption that not all parts of a document are equally important. In this way, the model is allowed to distinguish the importance of the different parts that compose a document, resulting in benefits in Sentiment Analysis.

Another studied family of neural networks is Convolutional Neural Network (CNN) architectures. CNN models were initially proposed in the field of computer vision (Lecun et al., 1998), although they have been successfully applied to a range of NLP tasks (Collobert et al., 2011). A representative study of the use of CNN models in Sentiment Analysis is

described in Kim (2014).

This work uses pre-trained vectors as well as embeddings that are allowed to be fine-tuned during training, which improves the model. Similarly, the work in Severyn et al. (2015) shows that the parameter initialization technique in the embeddings of a CNN model profoundly affects the final result. Also, this work incorporated noisy annotations from data to further refine the weights of the model, optimizing it to perform better in the task at hand. In Wehrmann et al. (2017), the authors propose a character-based CNN model that can be trained for several languages, without requiring any machine processes, comparing these results with the more traditional word-based encoding.

Taking this into consideration, it seems natural that some authors have studied the applicability of hybrid architectures, with the aim of leveraging both RNN and CNN models. In T. Chen et al. (2017), the proposed model aims at combining these two types of architectures into a unified approach that firstly classifies documents according to the number of opinion targets with a recurrent model and later applies a convolutional network to perform Sentiment Analysis. The work in Duyu Tang et al. (2015) tackles the encoding of semantic relations between sentences in a document by combining, on the one hand, a convolutional network that learns sentence representations, and on the other hand, a recurrent network that encodes the sentence relations.

Following, a new neural architecture has been proposed: the *Transformer* (Vaswani et al., 2017). The Transformer, in the context of NLP, is a relatively novel architecture that aims to solve sequence-to-sequence tasks while handling long-range dependencies. The main characteristic of the Transformer is that it is based solely on attention mechanisms, without any CNN or RNN architectures. Such an architecture allows for significantly more parallelization than the previous state-of-the-art models, which rely heavily on RNN-based architectures. Additionally, the Transformer also improves the accuracy in important NLP tasks, such as machine translation, and Sentiment Analysis (Ambartsoumian et al., 2018).

Based on the advancements that the Transformer brought to the NLP community, Devlin et al. (2018) presented Bidirectional Encoder Representations from Transformers (BERT). This Transformer-based language model trains deep bidirectional representations from unlabeled text by jointly observing both the left and right contexts. The BERT model has revolutionized the field, penetrating into all tasks within NLP. As a relevant example, BERT has been successfully applied to Sentiment Analysis (Sun et al., 2019).

Although the described neural architectures have high representational power, there have been efforts towards adapting embeddings models to the task of Sentiment Analysis. A pioneer work that uses embeddings applied to Sentiment Analysis is that of Maas et al. (2011).

In this work, both semantic and sentiment knowledge is captured in the embeddings. Similarly, the work in D. Tang et al. (2016) proposes a model that generates semantic sentiment embeddings, capturing the context of sentiment together with co-occurrence information. In this way, generated sentiment vectors can be directly used for Sentiment Analysis tasks without feature engineering. Sentiment signals that are incorporated into word embeddings can come from different sources, as in the case of the work in Li et al. (2017), where the model includes sentiment information from both word and document level. An alternative method for generating sentiment embeddings is possible: refining a pre-trained embeddings model (Yu et al., 2017). This work adjusts the original word embeddings, allowing them to be closer to both semantic and sentiment-wise similar words; while at the same time moving them further away to dissimilar words also in the semantic and sentiment plane.

Word embedding models are a key part of this thesis and are used extensively in the following publications: **“Enhancing deep learning sentiment analysis with ensemble techniques in social applications”** (Section 3.2.5), **“A semantic similarity-based perspective of affect lexicons for sentiment analysis”** (Section 3.2.3), **“An Approach for Radicalization Detection Based on Emotion Signals and Semantic Similarity”** (Section 3.2.1), and **“MoralStrength: Exploiting a moral lexicon and embedding similarity for moral foundations prediction”** (Section 3.2.2).

1.2.3 Semantic Similarity

Corpus-based semantic similarity methods are based on word associations learned from large text collections following the distributional hypothesis (Peter D Turney, Pantel, et al., 2010). Two words are assumed to be similar if their surrounding contexts are similar, or if they appear together more frequently. The computation of corpus-based methods is based on statistics of word distributions through word co-occurrences. According to different computational models, there are count-based methods, e.g., Pointwise Mutual Information (Church et al., 1990) or Normalized Google Distance (Gligorov et al., 2007), and predictive methods, such as Word2Vec (Mikolov et al., 2013). Count-based methods count word co-occurrences and construct a word-word matrix, in which those co-occurrence statistics are directly applied with probabilistic models (Church et al., 1990), matrix factorization (Pennington et al., 2014) and dimension reduction (Levy et al., 2015). Predictive-based methods directly learn dense vectors by predicting a word from its surrounding context. In this thesis, we use the predictive-based word embedding tool Word2Vec (Mikolov et al., 2013) to learn dense vector representation of words, because it has been reported to have good performance in many applications (Baroni et al., 2014). As suggested by the Word2Vec authors (Mikolov et al., 2013), the Continuous Bag of Words (CBOW) model is more computationally efficient and

suitable for larger corpus than the skip-gram model.

Thus, the CBOW model is used to train word vectors in a neural network architecture, which consists of an input layer, a projection layer, and an output layer to predict a word given its surrounding words with a certain context window size. Formally, given a sequence of training words $\{w_1, w_2, \dots, w_T\}$, each word vector is trained to maximize the average log probability:

$$\frac{1}{T} \sum_{t=1}^T \sum_{-c \leq k \leq c, k \neq 0} \log p(w_t | w_{t-k}, \dots, w_{t+k}) \quad (1.1)$$

where k is the context window size and $p(w_t | w_{t-k}, \dots, w_{t+k})$ is the hierarchical softmax of the word vectors (Mikolov et al., 2013). Having the trained word vectors (the dimension is predefined empirically and we set 300 in our experiments), word similarity is computed using standard cosine similarity.

Although the training process relies on a neural network-based supervised prediction model, the real training results are the vector representation of words instead of the neural network prediction model. Because of such an idea, the training of word embedding is unsupervised and can be applied in various textual corpus without labeled datasets. Furthermore, due to the simple neural network architecture and the use of hierarchical softmax, Word2Vec is able to address large corpora, and the training is more efficient.

Semantic similarity models are used frequently in NLP, but their application to lexical resources has not been thoroughly studied. Extracting knowledge from a lexicon in a manner that is directly usable by a learning model represents an area that has not been explored. This same idea partly motivates the work described in “**A semantic similarity-based perspective of affect lexicons for sentiment analysis**” (Section 3.2.3), where a novel method for extracting features from a lexicon by means of a similarity model is described.

1.2.4 Affect Analysis of Online Radicalism

The main problems that have been addressed in previous research for the automatic processing of online radicalization are analysis, detection, and prediction of radicalization (Correa et al., 2013; Fernandez et al., 2018).

Online radicalization analysis aims at providing actionable information to improve Law Enforcement Agencies (LEA)’s decision making processes. According to Correa et al. (2013), analysis approaches fall into two main categories: network-based (leaders, communities, and topology characteristics) and content-based (authorship identification, stylometric analysis, website activities, affect analysis and usage).

The works focused on detection can be classified into web mining and text mining approaches (Correa et al., 2013). Web mining works aim at detecting radical online content using techniques such as focused crawling (S. Agarwal et al., 2015b), while text classification techniques aim at developing a binary classification model based on text features (Rowe et al., 2016) combined with other features such as social dynamics (Rowe et al., 2016; Fernandez et al., 2018).

Regarding detection, different types of analysis have been proposed for understanding better online radicalization, including network and content-based analysis. While network analysis takes into account the social interactions, a content-based analysis is focused on analyzing several aspects of radical texts, such as affects, topics, or stylistic features.

Finally, several works have addressed prediction. Ferrara et al. (2016) propose a machine learning framework for detecting extremist supporters, predicting extremist user content adoption (i.e., a retweet of extremist content) and predicting interactions with extremists (i.e., reply direct messages from extremists). The framework considers three types of features: user and activity, timing, and network. S. Agarwal et al. (2015a) presents a survey focused on two problems, the automatic identification of online radicalization (hate promoting content, users and hidden communities) and the prediction of civil unrest related events (protests, riots, public demonstrations). The survey concludes that most works have used spatiotemporal features as discriminatory ones for predicting events. López-Sánchez et al. (2018) propose a system for predicting radicalization risks. They propose to generate alarms based on the radicalization influence of monitored users and the emotional load of the received retweets.

Table 1.2: Summary of Literature Survey regarding the use of affective technologies for processing online radicalization signals. Types are A: Analysis, D: Detection, P: Prediction.

Type	Ref.	Data	Method	Affect Model
A	(Abbasi et al., 2007)	Forum posts	Linear regression	Intensity of hate and violence terms
A	(Birmingham et al., 2009)	YouTube Group	Sentiment analysis of most frequent terms	Sentiment polarity using SentiWordNet.

Table 1.2: Summary of Literature Survey regarding the use of affective technologies for processing online radicalization signals. Types are A: Analysis, D: Detection, P: Prediction.

A	(Rowe et al., 2016)	pro-ISIS Twitter users	Definition of windows per analyzing the interaction and an algorithm for calculation the adoption of a term based on lexical, sharing and interactions homophily	Sentiment analysis (polarity) using MPQA (Deng et al., 2015) and Ar-SenL (Badaro et al., 2014)
A	(Jurek et al., 2015)	Twitter	Lexicon based Sentiment analysis	Sentiment (polarity) using SentiWordNet (Baccianella et al., 2010)
A	(H. Chen, 2008)	Dark Web Forum postings	Support Vector Regression ensemble	Manual annotation of the intensity (0-1) of violence, anger, hate, racism and sentiment
A	(Vergani et al., 2015)	Dabiq radical magazine	Sentiment analysis based on LIWC (Linguistic Inquiry and Word Count) (Pennebaker et al., 2001)	Anger, anxiety and emotion tone from LIWC
A	(Chalothorn et al., 2013)	Forum posts	Threshold of negative words to calculate sentence polarity	Sentiment (polarity) using SentiWordNet (Baccianella et al., 2010)
A	(Ghajar-Khosravi et al., 2016)	ISIS fan girls tweets and control datasets	Lexicon based Sentiment Analysis toward some topics identified by keywords	Sentiment polarity
A	(Dewan et al., 2017)	Images and text from Facebook posts re- lated to Paris attack	Domain transfer learning with SentiBank Borth et al., 2013 for images and LIWC (Pennebaker et al., 2001) for texts	Sentiment polarity

Table 1.2: Summary of Literature Survey regarding the use of affective technologies for processing online radicalization signals. Types are A: Analysis, D: Detection, P: Prediction.

D	(Ashcroft et al., 2015)	pro and anti-ISIS tweets	Sentiment analysis (polarity) based on CoreNLP (Manning et al., 2014) that uses a Recursive Deep Model over a Sentiment Treebank.	Feature based classification using SVM, Naive Bayes and Adaboost
D	(S. Agarwal et al., 2015c)	Twitter	Single-class SVM and KNN	Binary model to indicate if a tweet includes terms that indicate a negative emotion
D	(Saif et al., 2017)	pro and anti-ISIS tweets	Lexicon based Sentiment Analysis	Polarity (sentiment): neutral, positive, negative using SentiStrength (Thelwall et al., 2012)
P	(López-Sánchez et al., 2018)	Twitter	Alerts based on user influence (followers, followed and messaged retweeted) and sentiment load of received mentions	Sentiment (polarity) using SentiStrength (Thelwall et al., 2012)

In this section, we provide a more detailed analysis of the use of affect technologies for analyzing online radicalization. As shown in Table 1.2, affect analysis has been applied in a wide array of domains, such as radical forums (Abbasi et al., 2007; H. Chen, 2008; Chalothorn et al., 2013); radical magazines (Vergani et al., 2015); and social networks, such as Twitter (Rowe et al., 2016; Ghajar-Khosravi et al., 2016; Jurek et al., 2015; Saif et al., 2017; Rowe et al., 2016), Facebook (Dewan et al., 2017) and YouTube (Bermingham et al., 2009).

Regarding the affect model, most works analyze the polarity (i.e., valence or sentiment) using Sentiment Analysis techniques (Bermingham et al., 2009; Jurek et al., 2015; Vergani

et al., 2015; Chalothorn et al., 2013; Ghajar-Khosravi et al., 2016; S. Agarwal et al., 2015c; Saif et al., 2017). Other works evaluate the intensity of some terms, such as hate or violence (Abbasi et al., 2007; Jurek et al., 2015). Finally, other works (H. Chen, 2008; Dewan et al., 2017) use LIWC’s categories for affective processes (Pennebaker et al., 2001): positive emotions, negative emotions, anxiety, anger, and sadness.

The advancements in Sentiment Analysis for radicalization studies have provided many insights. Abbasi et al. (2007) conclude that there is a linear relationship between the intensity scores for violence and hate affects, which is strong in Middle Eastern forums and weak in U.S. forums.

Another interesting aspect is the process of radicalization. Rowe et al. (2016) noticed when analyzing the radicalization process that users, before being activated, frequently discuss politics, using words like ‘Syria’ or ‘Egypt’ with negative sentiment. Once they are activated, they tend to use more religious words, and some words as ISIS are used with negative sentiment since they prefer the term ‘Islamic State’.

Some analyses have also considered the temporal evolution of the text with different purposes, such as the comparison of different radical blogs (H. Chen, 2008) or the language used in Dabiq radical magazine (Vergani et al., 2015).

Finally, another compelling aspect is how public opinion reacts when a terrorist attack has happened. Dewan et al. (2017) analyzed both the sentiment of images and texts posted on Facebook. They observed that the sentiment for textual posts was negative during the first few hours but gradually moved to the positive side over time. Contrarily, posted images reflected positive sentiment initially but moved towards negative after the first few hours.

There is not an agreement in current research regarding the importance of affect signals. Several studies (Chalothorn et al., 2013; Ashcroft et al., 2015; S. Agarwal et al., 2015c) highlight their importance; while in other experiments (Saif et al., 2017) sentiment features, which consist of both word unigrams and their sentiment, have no impact on the classification performance compared to using unigrams only.

We have seen that previous works tackle successfully incorporating sentiment information for radicalization detection. Although the literature indicates that emotions are connected to the radicalization process, computational approaches that insert Emotion Analysis methods are scarce. The work presented in **“An Approach for Radicalization Detection Based on Emotion Signals and Semantic Similarity”** (Section 3.2.1) effectively exploits emotion information, incorporating it into the radicalization detection process. These affect cues are combined with textual representations based on word-embeddings, achieving

high-performance results.

1.2.5 Moral Foundations Theory

Psychologists and social scientists have systematically analyzed text data to address their research questions. Back in the late '60s, the Harvard General Inquirer dictionary (Stone et al., 1966) was the precursor of Sentiment Analysis which was to become, together with opinion extraction, a core theme of natural language processing (NLP).

Ever since, scientists gradually increased complexity moving from simple techniques (e.g., the unsupervised and partially rule-based approach of Peter D Turney (2002)) to sophisticated methods that try to determine the context of words (e.g., Muhammad et al. (2016)). The topics addressed also became more challenging tackling notions such as irony and sarcasm (Sulis et al., 2016). With time, not only the methods became more sophisticated, but also the tasks became more ambitious. Extensive studies on linguistic markers of sentiment and affect (Strapparava and Mihalcea, 2008; Liu, 2015; Muhammad et al., 2016; Poria, Gelbukh, et al., 2014) paved the way to assess more complex constructs such as personality (Schwartz et al., 2013; Yarkoni, 2010) and human values (Boyd et al., 2015; J. Chen et al., 2014). Moral values are considered to be a higher-level construct with respect to personality traits, determining how and when dispositions and attitudes relate with our life stories and narratives (McAdams et al., 2006). Here we provide a brief literature review of the studies regarding moral values assessment from textual data. As in sentiment and personality analysis, also, in this case, pioneer works followed a dictionary-based approach, while the current state-of-the-art performance is based on deep learning.

The first vocabulary developed to assess the moral values from textual data was the Moral Foundations Dictionary (MFD) (Graham et al., 2009). It was used together with the Linguistic Inquiry and Word Count (LIWC) software (Tausczik et al., 2010) to estimate moral traits and to investigate differences in moral concerns between different cultural groups. Clifford et al. (2013) employed the MFD for performing manual text analysis of 12 years of coverage in the New York Times, focusing on political debate in the US. Teernstra et al. (2016) assessed the political debate regarding the “Grexit” from approximately 8,000 tweets. They compared the performance of using the raw data, bi-grams, and the MFD features in employing basic machine learning models, namely, Naive Bayes (NB) and Maximum Entropy (ME). They concluded that pure machine learning is preferable to dictionary approaches since it has similar prediction accuracy while using fewer assumptions. In this thesis, we follow a similar approach to Teernstra et al. (2016); however, we propose an expanded version of the MFD, including also the moral valence per lemma. Moreover,

we employ logistic regression models to infer moral values from uni-grams combined with lexical features.

Dehghani et al. (2014) examined the differences between liberal and conservative moral value systems using a hierarchical generative topic modeling technique, based on Latent Dirichlet Allocation (LDA) (Blei et al., 2003), to enable the unsupervised detection of topics in their corpus of liberal and conservative weblogs. They used small sets of words selected from the MFD as seeds to encourage the emergence of issues related to different moral concerns. They examined similarities and differences in how such matters are expressed between these groups. Consistently with findings in moral psychology, they demonstrate that there are significant differences in how liberals and conservatives construct their moral belief systems. Sagi et al. (2014a) employed the same framework to study moral rhetoric in text for a specific case study, the US Federal shutdown of 2013 (Sagi et al., 2014b), where they examined the role of morals in intra- and inter-community differences of political party retweets. In both works, they were based on the framework presented in Dehghani et al. (2014), where LDA was employed to create a co-occurrence matrix on which the similarity between the texts and the vectors representing the different MFT moral traits was computed. In a similar approach, Kaur et al. (2016) attempted to quantify the moral loadings of text, based on the Latent Semantic Analysis (LSA). They used a bag-of-words model, representing the entire corpus by a word-context matrix. Then they reduced its dimensionality obtaining low-dimensional word vectors, in which similar vectors represent similar meaning words. Our study is presenting a different approach since we do not use LSA representations, but rather pre-trained word embeddings models. Although pre-trained word embeddings do not contain domain-specific knowledge, they express language regularities encoded as offsets in the resulting vector space. The proposed representations based on **“A semantic similarity-based perspective of affect lexicons for sentiment analysis”** (Section 3.2.3) precisely exploit the similarity between the analyzed text and a selection of words with moral content.

More recently, Garten, Boghrati, et al. (2016) employed the MFD to detect moral rhetoric in general, and more specifically, shifts in long political speeches over time. Then, based on psychological dictionaries and semantic similarity to quantify the presence of moral sentiment around a given topic, Garten, Joe Hoover, et al. (2018), proposed the Distributed Dictionary Representations (DDR) method. Showing promising results, DDR was also employed by Joe Hoover et al. (2018) to detect moral values in charitable giving. Later on, Garten, Kennedy, et al. (2019) extended the method, incorporating demographic embeddings into the language representations. Our approach is based on an expanded version of the MFD, with evaluated manual annotations regarding the moral valence of each lemma, that can be incorporated in computational frameworks.

In a study more similar to the one performed in this thesis, attempting to predict moral values involved in Twitter posts, Lin et al. (2018) proposed a method that automatically acquires background knowledge to improve the moral value prediction, pointing out the difficulty of the task also for human experts. Based on the work of Lin et al. (2018) and Graham et al. (2009), Mooijman et al. (2018) predicted the moral sentiment of the tweets. Their model consists of three layers, an embedding (lookup) layer, a recurrent neural network with long short-term memory (Hochreiter et al., 1997), and an output layer. The first layer converts words in an input tweet to a sequence of pre-trained word embeddings. The LSTM layer processes these embeddings and outputs a fixed-sized vector that encodes critical information for moral value prediction. A vector representing the percentage of words that match each category in the Moral Foundations Dictionary (Graham et al., 2009) are concatenated with the LSTM feature vector. Our approach is differentiating to this one since we employ word embeddings to compute the similarity between words rather than directly feeding them in a neural network architecture. Along the same line, Rezapour et al. (2019) investigated the relationship between moral values and stance towards a series of social issues. Their findings show that enhancing the original MFD improves the prediction accuracy of morality in text. They expanded the original MFD and employed a series of machine learning classifiers (SVM, RF, and LSTM) predicting the moral traits. This study underlines the importance of expanding the MFD; we go one step further, introducing the notion of moral “strength” while showing how more abundant information is improving the overall accuracy of the models.

Computational approaches to moral values are highly scarce, and one of the few computational resources is the MFD, which is limited in coverage. A higher-quality lexicon for this domain can benefit research, and thus a contribution of the work presented in **“Moral-Strength: Exploiting a moral lexicon and embedding similarity for moral foundations prediction”** (Section 3.2.2) is the extended lexicon of moral lemmas with the respective moral valence. To explore the properties and full potentials of the lexicon, we suggest three different models of increasing complexity, demonstrating the value of this resource. The proposed approaches range from feature engineering methods to a system which employs word embeddings of semantic similarity based on the work presented in **“A semantic similarity-based perspective of affect lexicons for sentiment analysis”** (Section 3.2.3).

1.3 Hypotheses

In light of the described context and background of this thesis, we have formulated the following hypotheses, that guide the course of this work:

- **Hypothesis 1 (H_1)**. The combination of deep and surface textual features, including affect lexicons, enhances final Sentiment Analysis performance.
- **Hypothesis 2 (H_2)**. A method that exploits both distributional representations and affect lexicons through semantic similarity can be leveraged for Sentiment Analysis effectively.
- **Hypothesis 3 (H_3)**. A model implementing a semantic similarity-based exploitation of a lexicon for Sentiment Analysis can be transferred to other Natural Language Processing areas that show domain-specific traits (e.g., radicalization detection, moral value estimation).

1.4 Objectives

With the objective of testing the validity of the previous hypotheses, we propose the following objectives, which are oriented to satisfy the motivation of this thesis:

- **Objective 1 (O_1)**. Definition of a framework for the combination of surface and deep features, as well as the evaluation methodology.

Such a framework should cover different feature combination methods. Besides, a reference evaluation methodology that uses known datasets should be defined.

- **Objective 2 (O_2)**. Definition of a lexicon generation methodology based on the fundamental properties of neural networks and an evaluation procedure for affect lexicons.

The lexicon generation method should take advantage of the backpropagation algorithm, allowing affect annotations to be automatically generated.

- **Objective 3 (O_3)**. Definition of a model for the exploitation of semantic similarity in distributed representations.

The model should exploit the knowledge contained in both word embedding models and affect lexicons, offering a new perspective on how to extract features from text.

- **Objective 4 (O_4)**. Transference of a similarity-based model on lexicons to different Natural Language Processing domains, evaluating how specific domain knowledge can be inserted.

The domain adaptation method should be independent of the specific domain.

1.5 Document outline

The rest of the document is structured as follows. Chapter 2 describes the research methodology used to conduct and evaluate the results of this thesis, the phases followed, and guidelines on research methods. In Chapter 3, the full text of the publications that compose the thesis compendium is shown. Each paper is presented in a separate section, with a detailed table, and the full text of the paper as originally published. Chapter 4 presents an overview of the contributions made in this thesis. Following, the obtained results are detailed and related to the objectives to which they contribute. After that, the conclusions obtained during the development of this thesis are discussed to finish with some future work lines. Lastly, Appendix A contains a full list of the publications of this thesis.

Methodology

The scientific method defines a number of principles widely known, yet extremely difficult to apply correctly. Still, all researchers around the globe make daily efforts to adhere to these principles. This chapter presents the experimental methodology followed in this thesis.

“If we knew what we were doing, it would not be research”. Anonymous.

2.1 Introduction

Researchers use experiments to answer questions. An experiment can be characterized by the treatments and experimental units to be used, and the responses that are measured (Oehlert, 2010). The treatments are applied to the experimental units, culminating with the measure of one or more responses. Experiments help researchers answer questions, avoiding biases and errors that may appear in other mechanisms. Experiments allow for a direct comparison between different experimental units or treatments. More importantly, we are in control of experiments, which allows us to make more robust inferences about the nature of differences that can be seen in an experiment.

Treatments, units, and the method specify the experimental design. Although the analysis phase is not an intrinsic part of the experimental design, a wise experimenter will take into account the analysis when planning an experiment. In a general tone, a successful experimental design should avoid systematic errors, be precise and objective, allow for an estimation of error, and have broad validity. In practice, fulfilling all these requirements is not feasible, which may take an experimenter to adopt a compromise. For example, exploring a wide range of hyperparameters of a learning model can result in excessively long computation times. Instead, a reduced subset can be used in order to approximate to a suboptimal response.

This thesis is an experimental one. Experiments lie at the core of the work presented in this document. Through experimenting, a set of hypotheses has been heavily tested, which has allowed us to extract a series of interesting conclusions. Such a methodology follows, of course, the scientific method.

Given the domain the thesis addresses, many specific tools, methods, measures, and good practices have been needed. As an example, in the machine learning field, a number of conventions and methodologies are usually used, including this thesis. Also, and thanks to the advance of software engineering applied to both machine learning and Natural Language Processing, the experimental design, execution, and replication are easily done, allowing for an iterative workflow that facilitates the application of the scientific method.

We would like to highlight the importance of this last characteristic of experimentation: replication. Currently, many scientific fields find themselves in a replication crisis, aggravated by the increase in the overall number of yearly scientific publications. This thesis tries to overcome these limitations in such a context, outputting several software artifacts that allow for the easy replication of results, methods, and models. We believe that modern scientific works should consider the aspects mentioned above, and more importantly, in a field as

machine learning, which is intimately connected to software engineering.

The following sections describe the experimental methodology used during the development of this thesis. These sections stress requirements, recommendations, and good practices in the umbrella of machine learning experimentation.

2.2 Machine Learning Experimentation

In an effort to systematize the methodology in this thesis, we have followed a known proposal for machine learning-driven research (Alpaydin, 2014). This guideline defines a series of steps that, in reality, are the same for any type of experimentation (Montgomery, 2005). The overall discussion revolves around a fixed method for approaching a machine learning experiment, independently of the domain, the learning variant, and even the data used. As in this thesis, we have tackled several objectives and domains. This methodology has been applied repeatedly, varying and adapting for each case. In this way, we follow a series of steps:

1. **Definition of the aim of the study.** This first step aims to define the scope of the study, assessing what is the overall questions that, with our experiments, we try to answer. In some cases, we may be interested in comparing two or more learning approaches to a specific problem, or learning whether a particular approach is valid for a domain where it has not been studied before. In general, we usually have two or more learning systems, and using the data available, we want to order them in terms of error.
2. **Selection of the quality measure.** In this phase, a decision about the quality measure must be taken. While in other machine learning setting, a standard measure in the error rate is used. In sentiment and emotion analysis, a more frequent measure in the accuracy, precision, and recall. However, as the data user's complexity rises, a more informative measure is the F1-score, which summarizes the precision and recall into a single metric. This distinction is also helpful at the design phase since attending to a single metric facilitates the process of making decisions, rather than using several measures.
3. **Choice of factors.** Identifying the factors of the machine learning system depends on the aim of the study. For example, if the objective is to fix an algorithm and finding the best hyperparameters, then those are the factors. In this thesis, the objective is not usually to find the best hyperparameters for the designed algorithms, but rather comparing different learning approaches applied to a problem. In these cases, the

factors would be the different approaches themselves. Still, there are studies where some hyperparameter tuning is done. In those, we have to consider said parameters as factors also.

4. **Choice of experimental design.** In this step, some decisions must be made to divide the available data for train, development, and test. How this division is done is important, since using small datasets leads to high variance responses, with non-significant differences among them. Also, it is relevant to keep replication data to a minimum, as in over-sampling. Following recommendations in the literature, a train/development/test schema has been followed when the datasets allow it. When we have needed to resample the data, over-sampling has been avoided or at least compared against the alternative under-sampling, which does not generate “artificial” samples.
5. **Performing the experiment.** In this phase, many general software engineering recommendations are applicable. It is important to consider experiment replication for both future experiments and knowledge transference. For example, seeding the used random number generator and saving said seed are always good practices. One also needs to consider the effects of software aging. In fact, during this thesis’s development, much external software has been drastically changed, and previous data, results, and code need to be correctly adapted when re-used. Also, it is relevant to mention that comparisons should be unbiased. When comparing other methods to the one proposed in this thesis, all approaches should be investigated equally diligently.
6. **Statistical analysis of the data.** In this step, we need to analyze the obtained results so that whatever conclusion we get is not subjective or due to chance. In this way, we cast a hypothesis that must be answered by a series of statistical methods that receive as input the conditions of the experiments, as well as the results obtained. A significant example of this thesis is the Friedman text (Demšar, 2006), which is explained in more detail in Section 2.4.
7. **Conclusions.** Once all data is collected and analyzed, we are in the position of drawing objective conclusions. Although this constitutes the final step, a frequent conclusion is a need for further experimentation. Consistently with the literature, in this thesis, most of the designs and experiments are iterative. Thus, we establish a feedback link between the design phases and the experimentation analysis, where the last one provides feedback for the tuning of the first.

2.3 Evaluation

Section 2.2 describes, in a general manner, the phases that we follow in this thesis for the development, experimentation, and evaluation of machine learning systems. Such phases are applied widely in the community, and can even be used in experiments that fall outside the machine learning domain. In this section, we describe standard metrics that are used for evaluating a machine learning system. This serves as a reference for the contents of the rest of this document, as these metrics have been used extensively during this thesis.

To evaluate the prediction of a machine learning model, authors usually compare the available labels with the model output, considering the following matrix:

	Predicted class	
True Class	Positive	Negative
Positive	tp : true positive	fn : false negative
Negative	fp : false positive	tn : true negative
Total	p'	n'

Table 2.1: Confusion matrix for a two-class scenario.

These four cases comprehend all possible cases of wrong and right answers of a given machine learning model, which are:

- *True positives* (tp) are the documents correctly classified as positive.
- *True Negatives* (tn) are the negative documents that have been correctly classified as negative.
- *Type I errors* (or false positives, fp) are negative documents that have been incorrectly classified as positives.
- *Type II errors* (or false negatives, fn) are positive documents that have been incorrectly classified as negative.

Once this is defined, computing the usual evaluation metrics is straightforward. The first metric to consider is the *accuracy*, which encodes the fraction of correct predictions of a machine learning system:

$$\text{Accuracy} = \frac{tp + tn}{N} \quad (2.1)$$

Although accuracy is a widely known metric, it does not serve as a reliable measure when using skewed datasets. This is due to the fact that, if the distribution among classes is not even, the accuracy metric can result high, which could take the evaluation to draw wrong conclusions. Consequently, alternative measures can be used to consider uneven data classes.

In this context, the *precision* can be defined as the number of correct positive results divided by the number of all positive results.

$$\text{Precision} = \frac{tp}{tp + fp} \quad (2.2)$$

Similarly, the *recall* corresponds to the number of correct positive results divided by the number of positive results that should have been predicted.

$$\text{Recall} = \frac{tp}{tp + fn} \quad (2.3)$$

The *F1-score* can be interpreted as a weighted average of these two values, reaching its best score at 1, and worst at 0. The expression for the F1-score is:

$$F_1 = \frac{2}{\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}} = 2 \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (2.4)$$

being P and R the precision and recall, respectively. In this way, the F-Score combines the precision and recall values into a single value, giving information on the performance of a classification system in a numerical piece of information.

2.4 Statistical Analysis

Once the performance metrics have been computed, some conclusions could be drawn. Nevertheless, performing a statistical analysis is a good practice, and should be done when possible (Alpaydin, 2014), avoiding subjective observations of the results. In this sense, during the development of this thesis, we have employed a number of statistical tools, with the aim of analyzing the evaluation results.

Among the used methods, it is worth to highlight the Friedman’s test (Demšar, 2006), which implements a comparison among different learning models based on their performance results on a number of datasets. Friedman’s test initially computes each learning model’s rank in each dataset, taking into account that the best performing method is assigned the lower ranking. An average of their ranks resolves ties. In this way, r_j^i is the rank of the j -th

of the k algorithms on the i -th of a total of N datasets. Friedman's test uses the comparison of average ranks:

$$R_j = \frac{1}{N} \sum_i r_j^i \quad (2.5)$$

With this average, the test states that the null-hypothesis is that all the algorithms are equal, so their ranks R_j are equal. Considering this, the Friedman statistic, with $k - 1$ degrees of freedom can be computed as:

$$\chi_F^2 = \frac{12N}{k(k+1)} \left(\sum_j R_j^2 - \frac{k(k+1)^2}{4} \right) \quad (2.6)$$

Nonetheless, it is shown that there is a more useful statistic that is distributed according to the F-distribution, and has $k - 1$ and $(k - 1)(N - 1)$ degrees of freedom Demšar, 2006. This statistic is referred to as the Friedman F, and is expressed as:

$$F_F = \frac{(N - 1)\chi_F^2}{N(k - 1) - \chi_F^2} \quad (2.7)$$

If the null-hypothesis of the Friedman test is rejected, post-hoc tests can be conducted, complementing the statistical analysis. In this thesis, we have used the Bonferroni-Dunn test, which allows us to compare the results of several learning models to a baseline. This post-hoc test can be computed by comparing the critical difference (CD) with a series of pre-computed critical values (Demšar, 2006). The CD is computed as:

$$CD = q_\alpha \sqrt{\frac{k(k+1)}{6N}} \quad (2.8)$$

Considering this, in order to assess whether a particular model improves over a baseline, the difference in average ranking is computed and compared against the CD. If the difference between the two models is greater than the CD, the null-hypothesis is rejected, and it can be said that the model improves over the baseline with a statistical significance.

Publications

This chapter presents the central papers published as part of this Ph.D. thesis. As required, not all the publications of this thesis are included in this chapter. A full summary of publications, including those not directly related to the objectives of this thesis, is included in Appendix A.

3.1 Introduction

The following tables describe the main publications of this Ph.D. thesis. Since this thesis is presented as a compendium of publications, this section includes publications that fulfill the requirements. For a full list of the publications that are related to the thesis, please see Appendix A. The following color scheme has been used: gray for conference papers, and

blue for journal papers.

Table 3.1: Publications that are directly related to this thesis.

Page	Title	Year	Venue	Ranking
38	An Approach for Radicalization Detection Based on Emotion Signals and Semantic Similarity	2020	IEEE Access	JCR 2019 Q1 (3.745)
55	MoralStrength: Exploiting a moral lexicon and embedding similarity for moral foundations prediction	2019	Knowledge-Based Systems	JCR 2019 Q1 (5.921)
67	A Semantic Similarity-based Perspective of Affect Lexicons for Sentiment Analysis	2019	Knowledge-Based Systems	JJCR 2019 Q1 (5.921)
82	DepecheMood++: a Bilingual Emotion Lexicon Built Through Simple Yet Powerful Techniques	2019	IEEE Transactions on Affective Computing	JCR 2019 Q1 (7.512)
95	Enhancing Deep Learning Sentiment Analysis with Ensemble Techniques in Social Applications	2017	Expert Systems with Applications	JCR 2019 Q1 (5.452)

3.2 List of central publications

3.2.1 An Approach for Radicalization Detection Based on Emotion Signals and Semantic Similarity