COMPOSITIONAL LANGUAGE PROCESSING FOR MULTILINGUAL SENTIMENT ANALYSIS

DAVID VILARES CALVO

Doctoral Thesis

2017

Compositional language processing for multilingual sentiment analysis

DAVID VILARES

Doctoral Thesis /2017

Advisors:

MIGUEL A. ALONSO
CARLOS GÓMEZ-RODRÍGUEZ

Ph.D. degree in Computer Science

UNIVERSIDADE DA CORUÑA

MIGUEL ÁNGEL ALONSO PARDO, Associate Professor at Departamento de Computación at Universidade da Coruña,

CARLOS GÓMEZ-RODRÍGUEZ, Associate Professor at Departamento de Computación at Universidade da Coruña,

hereby certify

that the dissertation entitled *Compositional Language Processing for Multilingual Sentiment Analysis*, submitted to Universidade da Coruña by DAVID VILARES CALVO, has been carried out under our supervision and meets all the requirements for the award of *Ph. D. degree with International Mention*.

MIGUEL ÁNGEL ALONSO PARDO
Main advisor

CARLOS GÓMEZ-RODRÍGUEZ
Second advisor

Opinion is the medium between knowledge and ignorance.

— Plato

ACKNOWLEDGMENTS

It has been a long trip, with hills and valleys, but in all of them I was lucky enough to meet people that have contributed to this thesis, and to whom I would like to show my gratitude.

Miguel A. Alonso and Carlos Gómez-Rodríguez were both bright mentors. Miguel introduced me to natural language processing and science, and gave me to read Pang and Lee (2008), which was later the inception of the topic of this thesis. His commitment, dedication, and always fast and accurate feedback have made possible this dissertation coming to an end. He has been an unbelievable advisor. With Carlos I have shared uncountable avid discussions from which I always could learn many interesting things about natural language processing, research and many other subjects of life. He is one of the most talented NLP scientists I know, and I consider myself very fortunate to have been advised by him during these years.

I am also grateful to the members of the LYS group for their technical and administrative support, and specially to Yerai Doval, for creating an enjoyable working atmosphere at Lab 0.1.

A very special mention is reserved to Mike Thelwall, Erik Cambria and Yulan He, which hosted and advised me during my stays at foreign universities. They spent time with me sharing their knowledge and led me into new fields of research, helping me grow up professionally, for which I always will be grateful.

On a personal note, the biggest thank you of this thesis goes to my beloved ones. To my parents, for fully supporting me and make me understand the importance of studying and working hard since I started school. To Silvia, who always is there where I need it. To Suso and Mónica, for keeping the door of their home open for me. And to my grandparents, for sharing their data as nobody else do, giving priceless advises.

Also, friends make difficult moments fly away fast. Thank you for being there to share a drink, a meal or simply a good conversation to make me forget about Ph. D. (and other) problems.

Last, this research was possible thanks to the economical support of: Plan I2C of Xunta de Galicia, MECD (Grant FPU13/01180), Inditex-UDC 2014 & 2016 grants for visits, FPU 2015 grants for short visits, MINECO and FEDER (Grants TIN2010-18552-C03-02, FFI2014-51978-C2-02) and Xunta de Galicia (Grants CN2012/008, CN2012/319, R2014/034).

ABSTRACT

This dissertation presents new approaches in the field of *sentiment analysis* and *polarity classification*, oriented towards obtaining the sentiment of a phrase, sentence or document from a natural language processing point of view. It makes a special emphasis on methods to handle semantic composionality, i. e. the ability to compound the sentiment of multiword phrases, where the global sentiment might be different or even opposite to the one coming from each of their their individual components; and the application of these methods to multilingual scenarios.

On the one hand, we introduce knowledge-based approaches to calculate the semantic orientation at the sentence level, that can handle different phenomena for the purpose at hand (e. g. negation, intensification or adversative subordinate clauses).

On the other hand, we describe how to build machine learning models to perform polarity classification from a different perspective, combining linguistic (lexical, syntactic and semantic) knowledge, with an emphasis in noisy and micro-texts.

Experiments on standard corpora and international evaluation campaigns show the competitiveness of the methods here proposed, in monolingual, multilingual and code-switching scenarios.

The contributions presented in the thesis have potential applications in the era of the Web 2.0 and social media, such as being able to determine what is the view of society about products, celebrities or events, identify their strengths and weaknesses or monitor how these opinions evolve over time. We also show how some of the proposed models can be useful for other data analysis tasks.

RESUMEN

Esta tesis presenta nuevas técnicas en el ámbito del *análisis del sentimiento* y la *clasificación de polaridad*, centradas en obtener el sentimiento de una frase, oración o documento siguiendo enfoques basados en procesamiento del lenguaje natural. En concreto, nos centramos en desarrollar métodos capaces de manejar la semántica composicional, es decir, con la capacidad de componer el sentimiento de oraciones donde la polaridad global puede ser distinta, o incluso opuesta, de la que se obtendría individualmente para cada uno de sus términos; y cómo dichos métodos pueden ser aplicados en entornos multilingües.

En la primera parte de este trabajo, introducimos aproximaciones basadas en conocimiento para calcular la orientación semántica a nivel de oración, teniendo en cuenta construcciones lingüísticas relevantes en el ámbito que nos ocupa (por ejemplo, la negación, intensificación, o las oraciones subordinadas adversativas).

En la segunda parte, describimos cómo construir clasificadores de polaridad basados en aprendizaje automático que combinan información léxica, sintáctica y semántica; centrándonos en su aplicación sobre textos cortos y de pobre calidad gramatical.

Los experimentos realizados sobre colecciones estándar y competiciones de evaluación internacionales muestran la efectividad de los métodos aquí propuestos en entornos monolingües, multilingües y de *code-switching*.

Las contribuciones presentadas en esta tesis tienen diversas aplicaciones en la era de la Web 2.0 y las redes sociales, como determinar la opinión que la sociedad tiene sobre un producto, celebridad o evento; identificar sus puntos fuertes y débiles o monitorizar cómo estas opiniones evolucionan a lo largo del tiempo. Por último, también mostramos cómo algunos de los modelos propuestos pueden ser útiles para otras tareas de análisis de datos.

RESUMO

Esta tese presenta novas técnicas no ámbito da *análise do sentimento* e da *clasificación da polaridade*, orientadas a obter o sentimento dunha frase, oración ou documento seguindo aproximacións baseadas no procesamento da linguaxe natural. En particular, centrámosnos en métodos capaces de manexar a semántica composicional: métodos coa habilidade para compor o sentimento de oracións onde o sentimento global pode ser distinto, ou incluso oposto, do que se obtería individualmente para cada un dos seus términos; e como ditos métodos poden ser aplicados en entornos multilingües.

Na primeira parte da tese, introducimos aproximacións baseadas en coñecemento; para calcular a orientación semántica a nivel de oración, tendo en conta construccións lingüísticas importantes no ámbito que nos ocupa (por exemplo, a negación, a intensificación ou as oracións subordinadas adversativas).

Na segunda parte, describimos como podemos construir clasificadores de polaridade baseados en aprendizaxe automática e que combinan información léxica, sintáctica e semántica, centrándonos en textos curtos e de pobre calidade gramatical.

Os experimentos levados a cabo sobre coleccións estándar e competicións de avaliación internacionais mostran a efectividade dos métodos aquí propostos, en entornos monolingües, multilingües e de *code-switching*.

As contribucións presentadas nesta tese teñen diversas aplicacións na era da Web 2.0 e das redes sociais, como determinar a opinión que a sociedade ten sobre un produto, celebridade ou evento; identificar os seus puntos fortes e febles ou monitorizar como esas opinións evolucionan o largo do tempo. Como punto final, tamén amosamos como algúns dos modelos aquí propostos poden ser útiles para outras tarefas de análise de datos.

ACRONYMS

CNN Convolutional Neural Network

LAS Labeled Attachment Score

ML Machine Learning

NLP Natural Language Processing

P Precision

POS Part-of-speech

R Recall

SA Sentiment Analysis

SO Semantic Orientation

SVM Support Vector Machines

TF-IDF Term Frequency - Inverse Document Frequency

UAS Unlabeled Attachment Score

CONTENTS

LIST OF FIGURES

LIST OF TABLES

Part I

INTRODUCTION AND PRELIMINARIES

1 INTRODUCTION

This dissertation presents different approaches to address *polarity classification* for *sentiment analysis* (Pang and Lee, 2008), i.e. determining if the sentiment of a piece of text (it could be a phrase, a sentence or a whole document) reflects a positive, negative or neutral opinion.

First, we introduce sentiment analysis and also present natural language processing techniques and resources that will be used throughout the chapters of this book.

Second, we present different knowledge-based methods to calculate the semantic orientation at the sentence level, proposing models that can handle phenomena such as intensification or negation, among others.

Third, we describe how we build machine-learning methods for the purpose at hand, combining linguistic (lexical, syntactic and semantic) features, which work under monolingual, multilingual and code-switching environments.

Finally, the fourth part of this dissertation contains additional research results, obtained by some of the proposed methods, when they are applied to competitive evaluation campaigns, additional corpora and real time analysis tasks related not just to sentiment analysis, but also to data analysis in general.

1.1 MOTIVATION

Analyzing and comprehending subjective information expressed in social media by users of different countries, cultures and ages has become a key asset in order to monitor public opinion about products, events or public figures. This is observable in people's day-to-day life. From film forums such as FilmAffinity, it is possible to see what viewers think about different aspects of a movie, and make a decision about what to watch based on their personal preferences. From travel forums such as TripAdvisor, it is feasible to find out a large number of views about the accommodation where someone is planning to spend their next vacations. From modern social networks such as Facebook, Twitter or Instagram, we can infer from the comments of users what is their point of view with respect to news, trends or pictures, among others. From all of them, humans can process information and transform it into knowledge to answer questions like: *'How is the acting of Edward Norton in American History X?'*, *'I mostly care about bed comfort and cleanliness, should I book this room?'* or *'How is people's opinion evolving about Samsung Electronics after the Samsung Galaxy Note 7 issue?'*.

Before the appearance of the Web 2.0, a common solution for obtaining information about questions like these and many others was using surveys and polls. However, these strategies were typically expensive, had a limited scope and were only valid for a short period of time. Currently, social media could provide an effective way to poll users (Wang et al., 2012), plan business strategies (Li and Li, 2013) and make marketing decisions (Bae and Lee, 2012). However, human monitoring of web reviews and opinions presents important obstacles. The vast amount of opinions expressed every day in blogs, forums or social media makes manual observation unfeasible. Moreover, the application of corpus-based techniques to extract good sentiment features presents some advantages with respect to relying on intuitions, such as exhaustiveness of the resulting list of subjective words and the capacity to capture implicitly subjective constructions (Pang, Lee, and Vaithyanathan, 2002).

In this context, *sentiment analysis* (SA) has arisen as a research field of *natural language processing* (NLP) that deals with the automatic analysis of subjective content. Many subtasks can be located within this field of research. The primary one consists in classifying the *sentiment* or *polarity* of a text as positive or negative, although it is also common to include additional categories to distinguish purely informative texts, and to differentiate the strength of the opinions.

This dissertation focuses on how to obtain the sentiment of a phrase, sentence or document from a computational linguistic point of view. In particular, we make a special emphasis on methods to handle *semantic composionality*, i.e. the ability to compound the sentiment of multiword phrases, where their global sentiment might be different or even opposite to the one coming from their individual components. For example, in the sentence *'He is not very handsome, but he has something that I really like'*, we want to design algorithms with the ability to infer that the word *'very'* emphasizes *'handsome'*, *'not'* affects the whole expression *'very handsome'*, and *'but'* decreases the relevance of *'He is not very handsome'* and increases the one of *'he has something that I really like'*. We also pay special attention to adapting these methods to multilingual environments.

1.2 BACKGROUND

1.2.1 *Early history*

Francis Ysidro Edgeworth hypothesized in 1881 the *hedonometer*, a machine with the necessary psychological capacities to continuously monitor the level of happiness or pleasure of an individual. This is one of the earliest formal mentions to the automatic computation of feelings and emotions shared by people. Even before, Bentham (1789) posed that the pleasure caused by an action could be estimated by

an algorithm, latter referred to as *hedonic calculus*, based on variables such as intensity or duration; and how this could be used, for example, with legislative purposes.

From Edgeworth's hedonometer initial idea, few scientists tried to formalize and address this problem from a computational linguistic point of view during the 20th century, as detailed in Pang and Lee (2008).[1] Carbonell (1979) posed in his thesis that modeling understanding of natural languages requires a model of the processes underlying human thought. He proposed a theory of subjective language understanding and implemented it on the domain of politics, developing a system that can model liberal and conservative ideological reasoning in natural language. Wilks and Bien (1983) described how to model a structure to keep not just an individual's beliefs about aspects of the world, but also about beliefs that other individuals might have; about those and other aspects of the world. Already in the 90's some authors started to other problems more related to the current definition of sentiment analysis. For example, Wiebe, Bruce, and O'Hara (1999) described how to develop a gold-standard dataset for subjective classification and how to handle disagrements between annotators, a common problem in sentiment analysis, due to different personal beliefs that different people might have; and Wiebe and Bruce (2001) described work in developing a probabilistic model able to split a text into segments, where each segment is composed of a number of subjective sentences that share the same point of view and come from the same agent.

1.2.2 *The rise of sentiment analysis*

The rise of sentiment analysis came in the early 21st century, partially pushed by the success of the Web 2.0 and the first social media, that allowed users to generate and share their own content easily. Turney (2002) introduced an unsupervised algorithm to compute the semantic orientation[2] (SO) of texts, based on a dictionary approach. Such polarity is computed as a combination of the SO's of the individual words. To calculate the SO of adjective and adverb phrases, the author uses PMI-IR[3] (Turney, 2001), which measures the mutual information of a phrase with respect to the words *'excellent'* and *'poor'*. Alternatively, Pang, Lee, and Vaithyanathan (2002) introduced how to train supervised learning SA models, e.g. naïve Bayes, Maximum Entropy classifiers or Support Vector Machines (SVM), using features such as unigrams, bigrams or part-of-speech tags.

1 Check Pang and Lee (2008) §1.4 for a more exhaustive bibliography of early sentiment analysis.

2 Semantic orientation: A value indicating the positive or negative strength of a word. Usually positive values represent positive words and negative ones indicate negative words.

3 PMI-IR: Pointwise Mutual Information and Information Retrieval.

These two complementary approaches are often used to illustrate the two main angles to tackle polarity classification: the lexicon-based and machine learning perspectives, respectively. The latter has been widely explored in the recent years, starting from bag-of-words models. Gamon (2004) evaluated the role of linguistic features such as part-of-speech tag tri-grams and constituent structure of phrases in sentiment classification. Empirical results showed that, although features of this kind obtain a low performance by themselves, they contribute positively to accuracy when they are included in word n-gram models. As labeling data might be expensive, Pak and Paroubek (2010) explored how to crawl Twitter using happy and sad emoticons and build an automatically labeled corpus based on them. They also illustrated how some part-of-speech tags are prone to be used in certain type of texts (e. g. they point out that superlative adverbs and possessive endings may indicate positive opinions). Paltoglou and Thelwall (2010) explored the use of information retrieval weighting schemes for sentiment analysis such as the classic *tf-idf* (term frequency - inverse document frequency) (Salton and McGill, 1986)[4], evaluating them in different data sets. Martínez Cámara et al. (2011) evaluated the performance of an SVM and a naïve Bayes classifier on classifying the polarity of a film corpus using different weight schemes (e. g. binary occurrence or TF-IDF). It is also common to include some linguistic-related processing for preparing features (Bakliwal et al., 2012; Montejo-Ráez et al., 2012), such as lemmatization, stemming or stop word removal.

With respect to lexicon-based models, early efforts focused on the development of subjectivity resources. In addition to Turney (2002), Kanayama and Nasukawa (2006) proposed an additional unsupervised method for the detection of polar clauses on domain-oriented sentiment analysis. Esuli and Sebastiani (2006) released SentiWordNet, a lexical resource where each synset from Wordnet (Miller, 1995) is assigned three scores (s, p) with $p \in \{positive, negative, neutral\}$ and $s \in [0.0, 1.0]$.

1.2.3 *Compositional sentiment analysis*

Although popular, traditional bag-of-words models and subjectivity lexica on its own cannot correctly interpret the semantic compositionality of multiword phrases, at least not in a scalable way. In this respect, approaches tackling this challenge have gained attention from the NLP research community.

4 In some way, weighting schemes such as *tf-idf* can be considered as a way of taking the context into account.

Composition in machine learning SA systems

As introduced above, a naïve approach to emulate the comprehension of the meaning of multiword phrases for SA consists in using n-grams of words, with $n > 1$ (this was already pointed out in Pang, Lee, and Vaithyanathan, 2002). This approach is limited by the curse of dimensionality, although crawling data from the target domain can help to reduce that problem (Kiritchenko, Zhu, and Mohammad, 2014). In recent years, different approaches have tackled this problem in different ways.

Joshi and Penstein-Rosé (2009) proposed generalized dependency triplets as features for subjectivity detection, capturing non-local relations. Greene and Resnik (2009) introduced observable proxies for underlying semantics to approximate the relevant semantic properties automatically as features in a supervised learning setting, on the basis that the connection between structure and implicit sentiment is mediated by semantic properties characterizing the interface between syntax and lexical semantics. However, their experiments are not directly comparable to conventional labeling for opinionated tests. Wu et al. (2009) defined a phrase-based dependency parsing approach and proposed a tree-kernel based SVM as a model for polarity classification. Nakagawa, Inui, and Kurohashi (2010) also employed dependency trees for sentiment classification, representing the polarity of each dependency subtree by a hidden variable and performing sentiment classification by means of Conditional Random Fields to finally compute the polarity of the whole sentence. Socher et al. (2012) modeled a recursive neural network that learns compositional vector representations for phrases and sentences of arbitrary syntactic type and length. Socher et al. (2013) presented an improved recursive deep model for SA over dependency trees, and trained it on a sentiment treebank tagged using Amazon Mechanical Turk, pushing the state of the art up to 85.4% on the Pang and Lee (2005) dataset. Kalchbrenner, Grefenstette, and Blunsom (2014) showed how convolutional neural networks (CNN) can be used for semantic modeling of sentences. The model implicitly captures local and non-local relations without the need of a parse tree. It can be adapted for any language, as long as enough data is available. Severyn and Moschitti (2015) showed the effectiveness of a CNN in a SemEval sentiment analysis shared task (Nakov et al., 2016a), although crawling tens of millions of messages was first required to achieve state-of-the-art results. With a different purpose, Poria, Cambria, and Gelbukh (2016) presented a deep learning approach for aspect extraction in opinion mining, classifying the terms of a sentence as aspect or non-aspect. The system was then enriched with linguistic patterns specifically defined for aspect-detection tasks, which helps improve the overall performance and shows the utility of combining supervised and rule-based approaches.

In spite of being powerful and accurate, supervised approaches like these also present drawbacks. Firstly, they behave as a black box. Secondly, they do not perform so well on domain transfer applications (Aue and Gamon, 2005; Pang and Lee, 2008). Finally, feature and hyper-parameter engineering can be time and resource costly tasks.

Composition in knowledge-based SA systems

When the said limitations of machine learning models need to be addressed, unsupervised approaches are useful. Taboada et al. (2011) presented a lexical rule-based approach to handle relevant linguistic phenomena such as intensification, negation, *'but'* clauses and *irrealis*. Thelwall, Buckley, and Paltoglou (2012) released SentiStrength, a multilingual unsupervised system for micro-texts that handles negation and intensification, among other web linguistic phenomena. It is limited to window-based and word-matching rules, since no NLP phases such as part-of-speech tagging or parsing are applied. Regarding syntax-based approaches, the few described in the literature are language-dependent. Jia, Yu, and Meng (2009) defined a set of syntax-based rules for handling negation in English. Cambria, Olsher, and Dheeraj (2014) released SenticNet v3, a resource for performing sentiment analysis in English texts at the semantic level rather than at the syntactic level, by combining existing resources such as ConceptNet (Liu and Singh, 2004) and AffectiveSpace (Cambria et al., 2009). By exploiting artificial intelligence, semantic web technologies and dimensionality reduction techniques it computes the polarity of multiword common-sense concepts (e. g. *'buy Christmas present'*). With a different goal, Liu et al. (2016) automatically selected syntactic rules for an unsupervised aspect extraction approach, showing the utility of rule-based systems on opinion mining tasks.

1.2.4 *Multilingual sentiment analysis*

Monolingual sentiment analysis systems have been created for languages belonging to a variety of language families, such as Afro-Asiatic (Aldayel and Azmi, 2016), Indo-European (Ghorbel and Jacot, 2011; Habernal, Ptáček, and Steinberger, 2014; Medagoda, Shanmuganathan, and Whalley, 2013; Neri et al., 2012; Scholz and Conrad, 2013), Japonic (Arakawa et al., 2014), Sino-Tibetan (Vinodhini and Chandrasekaran, 2012; Zhang et al., 2009) and Tai-Kadai (Inrak and Sinthupinyo, 2010), among others.

The performance of a given approach for sentiment analysis varies from language to language. In the case of supervised systems, the size of the training set is a relevant factor (Cheng and Zhulyn, 2012; Demirtas and Pechenizkiy, 2013), but performance is also affected by linguistic particularities (Boiy and Moens, 2009; Wan, 2009) and the availability of language processing tools (Klinger and Cimiano,

2014) and resources (Severyn et al., 2016). With respect to the latter point, sentiment lexica are scarce for languages other than English, and therefore a great deal of effort has been dedicated to building lexical resources for sentiment analysis (Chen and Skiena, 2014; Cruz et al., 2014b; Gao et al., 2013; Hogenboom et al., 2014; Kim et al., 2009; Volkova, Wilson, and Yarowsky, 2013). A common approach for obtaining a lexicon for a new language consists in translating pre-existent English lexica (Brooke, Tofiloski, and Taboada, 2009), but it was found that even if the translation is correct, two parallel words do not always share the same semantic orientation across languages due to differences in common usage (Ghorbel and Jacot, 2011).

Another approach for building a monolingual SA system for a new language is based on the use of machine translation (MT) in order to translate the text into English automatically, to then apply a polarity classifier for English, yielding as a result a kind of cross-language sentiment analysis system (Balahur and Turchi, 2012b; Martínez Cámara et al., 2014; Perea-Ortega et al., 2013; Wan, 2009). It was found that text with more sentiment is harder to translate than text with less sentiment (Chen and Zhu, 2014) and that translation errors produce an increase in the sparseness of features, a fact that degrades performance (Balahur and Turchi, 2012a, 2014). To deal with this issue, several methods have been proposed to reduce translation errors, such as applying both directions of translation simultaneously (Hajmohammadi, Ibrahim, and Selamat, 2014) or enriching the MT system with sentiment patterns (Hiroshi, Tetsuya, and Hideo, 2004). In the case of supervised systems, self-training and co-training techniques have also been explored to improve performance (Gui et al., 2013, 2014).

Few multilingual systems for SA tasks have been described in the literature. Banea, Mihalcea, and Wiebe (2010) and Banea, Mihalceaa, and Wiebe (2014) described a system for detecting subjectivity (i. e. determining if a text contains subjective or objective information) in English and Romanian texts, finding that 90% of word senses maintained their subjectivity content across both languages. Xiao and Guo (2012) confirm on the same dataset that boosting on several languages improves performance for subjectivity classification with respect to monolingual methods.

Regarding the few multilingual polarity classification systems described in the literature, they are based on a supervised setting. In this respect, Yan et al. (2014) described a supervised multilingual system for SA working on previously tokenized Chinese and English texts. Some approaches rely on MT to deal with multi-linguality. Balahur et al. (2014) built a supervised multilingual SA system by translating the English SemEval 2013 Twitter dataset (Chowdhury et al., 2013) into other languages by means of MT, which improves on the results of monolingual systems due to the fact that, when multiple

languages are used to build the classifier, the features that are relevant are automatically selected.

Other approaches advocate the use of language-independent indicators of sentiment, such as emoticons (Davies and Ghahramani, 2011), for building language-independent SA systems, although the accuracy of a system built following this approach is worse than the combined accuracy of monolingual systems (Narr, Hülfenhaus, and Alnayrak, 2012). The use of other language-independent indicators, such as character and punctuation repetitions, results in low recall (Cui et al., 2011).

1.3 CONTRIBUTIONS

In this context, the work reported in this dissertation has contributed to the effective advancement of the state of the art in SA and related text-mining application by means of the formal definition of techniques for text processing, their implementation into practical tools and the construction of language resources.

In what follows, we list the main contributions of the thesis:

1. A set of pre-trained of bilingual and multilingual parsers that can be used to perform multilingual sentiment analysis.

2. A Spanish version for SentiStrength, a widely used multilingual lexicon-based system for SA, specially on micro-text scenarios (e. g. Twitter or Youtube) and a SentiStrength Spanish Twitter corpus annotated with sentiment information.

3. A Spanish syntax-based system for SA on long reviews. The system works for Spanish using trees annotated under Ancora (Taulé, Martí, and Recasens, 2008) guidelines.

4. A formalism for compositional operations, allowing the creation of arbitrarily complex rules to tackle relevant phenomena for SA, for any language and syntactic dependency annotation. We implement and evaluate a set of practical universal operations defined using part-of-speech (PoS) tags and dependency types under the universal guidelines of Petrov, Das, and McDonald (2011), McDonald et al. (2013) and Nivre et al. (2016): universal annotation criteria that can be used to represent the morphology and syntax of any language in a uniform way.

5. Novel methods for classifying the polarity of Spanish tweets by using linguistic knowledge. The main contribution consists in building models which combine lexical, syntactic, psychometric and semantic knowledge to illustrate the performance that linguistic perspectives can achieve, ranging from shallow to deep knowledge, on monolingual, multilingual and code-switching scenarios.

6. The first English-Spanish code-switching corpus for SA annotated according to the SentiStrength and trinary (positive, negative and neutral) scales.

7. A real time analysis of Spanish politicians based on what the public says about them in Twitter.

1.4 STRUCTURE OF THE THESIS

This dissertation is structured in four parts. This first part is introductory: it self-contains this chapter, where we also list the publications derived from our work. Additionally, it describes different NLP techniques used throughout the thesis. The second part presents knowledge-based approaches to handle semantic composionality at the sentence and document levels. The third part describes a machine-learning approach for polarity classification and shows its robustness on monolingual, multilingual and code-switching environments. The fourth and last part is devoted to the applications of the previous chapters in evaluation campaigns and other data analysis tasks.

A chapter-by-chapter breakdown of each of the parts follows:

PART I:

CHAPTER 2 introduces core NLP tasks, such as preprocessing, part-of-speech tagging and dependency parsing. Those represent the core techniques and algorithms that we will be using throughout this dissertation to tackle sentiment analysis.

PART II:

CHAPTER 3 presents Spanish SentiStrength, a Spanish version of the widely used lexicon-based method SentiStrength (Thelwall et al., 2010). Additionally, we introduce a Spanish sentiment corpus annotated according to the dual score used by SentiStrength (i. e. each text, receives both a positive and negative score) and explore the influence of different phenomena on determining the sentiment of tweets.

CHAPTER 4 describes a Spanish syntax-based system to determine the semantic orientation on long reviews coming from forums. We propose a prefixed set of syntactic rules applicable to trees annotated according to Ancora guidelines (Taulé, Martí, and Recasens, 2008), indicating its scope and desired behavior. We explore the impact of different rules and finally show how creating domain-dependent dictionaries significantly increases the performance on specific-dependent datasets.

CHAPTER 5 introduces a theoretical formalism that allows to create arbitrarily complex rules to tackle relevant phenomena for SA, for any language and syntactic dependency annotation. We implement and evaluate a set of practical universal operations defined using part-of-speech tags and dependency types under the Universal Treebanks (McDonald et al., 2013) and Universal Dependencies (Nivre et al., 2016). The system is compared against existing approaches on different languages.

PART III:

CHAPTER 6 proposes different linguistic features for supervised SA on Spanish tweets, including lexical, psychometric and syntactic features in the form of enriched generalized dependency triplets, and explores their impact on determining the polarity of Spanish tweets.

CHAPTER 7 applies the features and models described in Chapter 6 to monolingual, multilingual and code-witching settings, comparing them with language-dependent pipelines.

PART IV:

CHAPTER 8 applies the model proposed in Chapter 6 to a different task: multi-label topic classification, and draws which features are more relevant for the purpose at hand.

CHAPTER 9 applies Spanish SentiStrength to real-time analysis of political tweets. It shows how the sentiment expressed about politicians on Twitter is similar to the one reported by polls, but this does not translate to an accurate prediction of election results.

CHAPTER 10 summarizes the performance of some of the models described in previous chapters on RepLab 2014 (Amigó et al., 2014), an evaluation campaign to monitor reputation classification of entities and also to distinguish influential from non-infuential authors in Twitter.

CHAPTER 11 describes our participation at SemEval 2016 (Nakov et al., 2016a) and briefly introduces a deep learning model based on convolutional neural networks and the use of deep features.

PART V:

CHAPTER 12 summarizes the work presented in this thesis, draws our conclusion, and presents future lines of research.

1.5 PUBLICATIONS

Part of the research presented in this dissertation has been published in peer-reviewed conference and workshop proceedings, as well as indexed journals. A list of these publications follows:

JOURNALS:

- David Vilares, Carlos Gómez-Rodríguez, and Miguel A. Alonso (2017a). "Supervised Sentiment Analysis in Multilingual Environments." In: *Information Processing & Management* 53, pp. 595–607
- David Vilares, Carlos Gómez-Rodríguez, and Miguel A. Alonso (2017b). "Universal, unsupervised (rule-based), uncovered sentiment analysis." In: *Knowledge-Based Systems* 118, pp. 45–55
- David Vilares and Miguel Alonso (2016). "A review on political analysis and social media." In: *Procesamiento del Lenguaje Natural* 56, pp. 13–24
- David Vilares, Miguel A. Alonso, and Carlos Gómez-Rodríguez (2015b). "On the usefulness of lexical and syntactic processing in polarity classification of Twitter messages." In: *Journal of the Association for Information Science and Technology* 66.9, pp. 1799–1816
- David Vilares, Mike Thelwall, and Miguel A. Alonso (2015). "The megaphone of the people? Spanish SentiStrength for real-time analysis of political tweets." In: *Journal of Information Science* 41.6, pp. 799–813
- David Vilares, Miguel A. Alonso, and Carlos Gómez-Rodríguez (2015a). "A syntactic approach for opinion mining on Spanish reviews." In: *Natural Language Engineering* 21.01, pp. 139–163
- David Vilares, Miguel A. Alonso, and Carlos Gómez-Rodríiguez (2013). "Una aproximación supervisada para la minería de opiniones sobre tuits en español en base a conocimiento lingüístico." In: *Procesamiento del lenguaje natural* 51, pp. 127–134. ISSN: 1135-5948
- David Vilares, Miguel A. Alonso, and Carlos Gómez-Rodríguez (2013). "Clasificación de polaridad en textos con opiniones en español mediante análisis sintáctico de dependencias." In: *Procesamiento de Lenguaje Natural* 50, pp. 13–20

CONFERENCES AND WORKSHOPS

- David Vilares, Carlos Gómez-Rodríguez, and Miguel A. Alonso (2016). "One model, two languages: training bilingual parsers with harmonized treebanks." In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Berlin, Germany: Association for Computational Linguistics, pp. 425–431

- David Vilares, Yerai Doval, Miguel A. Alonso, and Carlos Gómez-Rodríguez (2016). "LyS at SemEval-2016 Task 4: Exploiting Neural Activation Values for Twitter Sentiment Classification and Quantification." In: *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*. San Diego, California: Association for Computational Linguistics, pp. 79–84

- David Vilares, Miguel A. Alonso, and Carlos Gómez-Rodríguez (2016). "EN-ES-CS: An English-Spanish Code-Switching Twitter Corpus for Multilingual Sentiment Analysis." In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. Portorož, Slovenia: European Language Resources Association (ELRA), pp. 4149–4153

- David Vilares, Miguel A. Alonso, and Carlos Gómez-Rodríguez (2015). "Sentiment Analysis on Monolingual, Multilingual and Code-Switching Twitter Corpora." In: *Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. Lisboa, Portugal: Association for Computational Linguistics, pp. 2–8

- David Vilares, Miguel Hermo, Miguel A. Alonso, Carlos Gómez-Rodríguez, and Yerai Doval (2014a). "LyS: Porting a Twitter Sentiment Analysis Approach from Spanish to English." In: *Proceedings og The 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pp. 411–415

- David Vilares, Miguel Hermo, Miguel A. Alonso, Carlos Gómez-Rodríguez, and Jesús Vilares (2014b). "LyS at CLEF RepLab 2014: Creating the state of the art in author influence ranking and reputation classification on Twitter." In: *Proceedings of the Fifth International Conference of the CLEF initiative*, pp. 1468–1478

- David Vilares, Yerai Doval, Miguel A. Alonso, and Carlos Gómez-Rodrıguez (2014c). "Lys at TASS 2014: A prototype for extracting and analysing aspects from Spanish tweets." In: *Proceedings of the TASS workshop at SEPLN*

- David Vilares, Miguel A. Alonso, and Carlos Gómez-Rodríguez (2013). "Supervised polarity classification of Spanish tweets based on linguistic knowledge." In: *DocEng'13. Proceedings of the 13th ACM Symposium on Document Engineering*. ACM. Florence, Italy, pp. 169–172

- David Vilares, Miguel A. Alonso, and Carlos Gómez-Rodríguez (2013). "LyS at TASS 2013: Analysing Spanish tweets by means of dependency parsing, semantic-oriented lexicons and psychometric word-properties." In: *Proc. of the TASS workshop at SEPLN 2013. IV Congreso Español de Informática*, pp. 179–186
- Miguel A. Alonso, Carlos Gómez-Rodríguez, David Vilares, Yerai Doval, and Jesús Vilares (2015). "Seguimiento y análisis automático de contenidos en redes sociales." In: *Actas: III Congreso Nacional de i+d en Defensa y Seguridad, DESEi+d 2015*. Marín, Spain: Centro Universitario de la Defensa de Marín, pp. 899–906

1.6 SOFTWARE AND RESOURCES

As results of this dissertation, different resources and software have been made available for the research community:

- Spanish SentiStrength can be tried at this website.[5]

  The needed data to make it work, can be downloaded from this link.[6] To obtain a research version of the SentiStrength code, contact either David Vilares (`david.vilares@udc.es`) or Mike Thelwall (`M.Thelwall@wlv.ac.uk`).

- Universal, Unsupervised (rule-based), Uncovered SA (UUUSA), corresponding to the implementation of the formalism introduced in Chapter 5, can be downloaded at this link.[7]

- Mineria de Opiniones mediante Inteligencia Artificial (MIOPIA) can be downloaded here[8], which includes functionalities to try the enriched generalized dependency triplets and lexical-based features explained in Chapter 6 and also includes a version of the system introduced in Chapter 4.

- A Spanish corpus annotated according to the SentiStrength criteria can be obtained here.[9]

- An English-Spanish code-switching corpus annotated according to the SentiStrength and trinary scale criteria can be downloaded from this link.[10]

- A set of pretrained bilingual parsers can be found at this website.[11]

5 http://sentistrength.wlv.ac.uk/#Non-English
6 http://sentistrength.wlv.ac.uk/SpanishSentiDataDavidVilares.zip
7 http://grupolys.org/software/UUUSA/
8 http://grupolys.org/software/MIOPIA/
9 http://sentistrength.wlv.ac.uk/SpanishTweetsTestAndDevelopmentSetsDavidVilares.zip
10 http://grupolys.org/software/CS-CORPORA/
11 https://github.com/aghie/pretrained-multilingual-parsers

2

PRELIMINARIES

In this chapter we introduce natural language processing techniques that will be used throughout this dissertation.

2.1 PREPROCESSING

Gargabe in, garbage out is a common problem in computer science, and in particular in data mining, that refers to the undesired outputs produced by the systems when they have to deal with input data of poor quality. In the context of sentiment analysis, this translates into users that write using typos, misspelled words or ungrammatical sentences when sharing their views, that will serve as the input to the systems. To counteract this, a preprocessing pipeline is applied to the data sets that we will be using in this dissertation. This includes:

• *Normalization of punctuation marks*. Not respecting punctuation rules in forums or social networks is a handicap for the rest of processing, especially tokenization. To resolve this, all punctuation mark representation will be modified by adding blanks when required. See Example 1.

EXAMPLE 1 (Normalization of punctuation marks). Given the sentence *'I like it,but it is too expensive'* the preprocessing pipeline will transform it into *'I like it, but it is too expensive'*.

□

• *Laughs normalization*. The irregular ways to express laughs are translated to *hxhx*, where $x \in \{a, e, i, o, u\}$, so as to be able to treat laughs in an unified way (see Example 2). A list of regular expressions is used to match the most common ways to simulate laughs in web texts. Interjections such as *'ha'* or *'hah'* are skipped, because it is hypothesized they do not represent actual laughs, being often part of sarcasms or complaints.

EXAMPLE 2 (Laughs normalization). Given the irregular laughs *'hhhaaha'* or *'hEhhhhE'*, the preprocessing pipeline will transform them into *'haha'* or *'hehe'*. In the case of other languages, such as Spanish, laughs are expressed differently (e. g. *'jajaja' 'hahaha'*), and it can be also handled by the pipeline.

□

• *Unification of compound expressions*. There are many compound expressions, that must usually be interpreted as single units of meaning.

To find them, a dictionary of compound expressions is used[1]. If the preprocessing pipeline identifies a group of these words, they are preprocessed as in Example 3.

EXAMPLE 3 (Unification of compound expressions). Given the Spanish compound expression *'sin embargo'* (*'however'*) or *'en absoluto'* (*'not at all'*), the pipeline will unify them into a single token (e. g. *'sin embargo'* becomes *'sin_embargo'* and *'en absoluto'* becomes *'en_absoluto'*).

□

In addition, when preprocessing Twitter, we must take into account the use of a very informal language combined with specific Twitter elements that do not appear in other social networks:

• *Emoticon preprocessing*. Emoticons are a combination of characters that somehow represent a human face sharing an emotion. In SA handling this phenomenon is relevant, especially at sentence-level polarity classification. In Chapters 4, 6 and 7 we deal with emoticons by preprocessing them, as indicated in Example 4, meanwhile in others we are using techniques (Gimpel et al., 2011; Thelwall, Buckley, and Paltoglou, 2012) that are able to keep them as a single unit of meaning after tokenizing the text (Chapters 3 or 5). In the former case, the preprocessing algorithm replaces the form of the emoticon by a string which represents the class, relying on emoticon lists (Agarwal et al., 2011) that distinguish five classes of emoticons: emoticon-strong-positive, emoticon-positive, emoticon-neutral, emoticon-negative, and emoticon-strong-negative. When they are not preprocessed, the list provided by SentiStrength (Thelwall, Buckley, and Paltoglou, 2012) is used to assign a semantic orientation to each emoticon, handling them as a single unit of meaning.

EXAMPLE 4 (Emoticon normalization). Given the sentence *'I am happy :)'*, the pipeline will transform it into *'I am happy. Emoticon-positive.'*, according to Agarwal et al. (2011) list, in order not to interfere with the subsequent tagging and parsing steps, when texts are tokenized using standard tools[2].

□

• *Treatment of Twitter special symbols ('@' and '#')*. The use of Twitter special symbols is an important issue, not only for text analytics, but also for segmentation and tokenization, as they can affect the performance of these processes. In the case of user mentions the '@' symbol is removed and the first letter is capitalized, because we hypothesize that user mentions usually refer to a proper name. An effective treatment of hashtags ('#') is more complex. A hashtag can be formed by a

1 In the case of Spanish those are extracted from the Ancora corpus (Taulé, Martí, and Recasens, 2008)

2 Stantard tokenization tools such as some of the ones provided by NLTK or Stanford CoreNLP, for example.

concatenation of multiple words, and often it refers to a very specific event and includes unknown words. In this case, a simple strategy can be followed: If the hashtag appears at the beginning or the end of the tweet, just remove it completely: we suppose that, in these cases, users only want to label their tweets. Otherwise, delete the '#', because we hypothesize that the rest of the hashtag contributes to syntactic information. Example 5 illustrates it.

EXAMPLE 5 (Normalization of Twitter nicknames and hashtags). Given the hypothetical tweet *'@user you are so #good man, #goodblessyou'* the pipeline will preprocess it as *'User you are so good man,'*. The pipeline used in this dissertation cannot properly handle composite hashtags, such as *'#word1_word2'* or *'#word1word2'*, which will be taken as a unique token during the whole preprocessing of the tweet.

□

- *URL normalization*: Web addresses that appear in tweets are identified and changed to a normalized string.

2.2 PART-OF-SPEECH TAGGING

Part-of-speech (PoS) tagging is the process of marking up a word in a text as corresponding to a part of speech, based on both its definition and its context. Part-of-speech tags can be coarse-grained (when they only represent the grammatical category: noun, verb, adjective, etc.) or fine-grained (when they include additional morpho-syntactic information such as gender, number, tense, etc.). Definition 1 introduces tagging from a formal point of view.

DEFINITION 1. Let $w=w_1,...,w_n$ be a sentence, where each word occurrence $w_i \in W$, with W being the vocabulary, and let $p_i \in P$ a part-of-speech tag indicating a grammatical category (e.g. noun, verb or adjective), a **tagged sentence** is a list of tuples (w_i, p_i) where each w_i is assigned a part-of-speech tag, p_i.

□

PoS tagging is a core process in natural language processing that has been addressed for decades (Brants, 2000; Brill, 1992; Giménez and Marquez, 2004; Ratnaparkhi et al., 1996; Søgaard, 2011; Toutanova and Manning, 2000). Example 6 illustrates the task with a valid part-of-speech-tagging output for a sentence. One of the main challenges of PoS tagging comes from homonyms, and in particular in written language from homographs, since the same word form can play different roles depending on the context, as shown in Example 7. Another common challenge involves classifying new and unknown words, that have not been seen in the dataset used to build the tagger. In both cases, context is crucial, and taggers try to learn such context in order to correctly predict the tag for the word at hand.

EXAMPLE 6 (A part-of-speech tagging of a sentence). A valid PoS-tagging output for the sentence *'He is not very handsome, but he has something that I really like'*.

He	**is**	**not**	**very**	**handsome**	**,**	**but**	**he**	**has**
PRON	VERB	ADV	ADV	ADJ	PUNCT	CONJ	PRON	VERB
something	**that**	**I**	**really**	**like**				
ADV	ADP	PRON	ADV	VERB				

The example uses the universal tagset proposed by Petrov, Das, and McDonald (2011). It includes tags (among others) for: pronoun (PRON), verb (VERB), adverb (ADV), adjective (ADJ), punctuation mark (PUNCT), conjunction (CONJ) and prepositions and postpositions (ADP).

□

EXAMPLE 7 (Part-of-speech tagging and homonym). Given the two sentences:

'Today there is a match between Barcelona and Real Madrid at 19:45 pm.' (*'match'* is a NOUN)

'These two colors match perfectly.' (*'match'* is a VERB)

The word *'match'* is a case of homonym, which only can be solved by looking at the neighbors, e. g. if the previous word is a determiner it is likely that *'match'* will be a noun, meanwhile if the previous word is a noun the probability of being a verb increases.

□

For informal texts, such as those used in social media, there are different approaches able to tackle their particularities, although mainly focused on English texts (Foster et al., 2011; Gimpel et al., 2011). For languages different from English, e. g. Spanish, PoS tagging in web environments must deal with absence of acute accents on sentences, that might have influence in the accuracy. To counteract this, a simple but effective solution consists in *expanding* the training set: each sentence is duplicated and all acute accents are removed from the copy. Although removing accents increases ambiguity, it keeps a state-of-the-art performance of the tagger and at the same time it is possible to handle homonyms based on neighbor words. This is shown in Example 8, where we can compare the output of a regular tagger and a tagger that learns from the expanded set.

EXAMPLE 8 (PoS tagging of a Spanish sentence with homonyms when acute accents are missing). We show the output provided by a Spanish tagger trained on the expanded Ancora training set[3] (that we will

3 For clarity reasons we are illustrating it using the universal tagset of Petrov, Das, and McDonald (2011), although the real output corresponds to Ancora tags (Taulé, Martí, and Recasens, 2008) and was first introduced in Vilares, Alonso, and Gómez-Rodríguez (2015a).

be using later in Chapter 4), compared to the one trained on the regular training set, when analyzing the sentence *'No he tenido tiempo de escribir sobre el y ya esta estropeado'*, which is not correctly written. The correct Spanish sentence would be *'No he tenido tiempo de escribir sobre él y ya está estropeado'* (*'I had no time to write about it and it is already broken'*). The issue is that Spanish language uses these diacritical accents to distinguish the meaning of *'el'* (*'the'*, determiner) from *'él'* (*'it'*, pronoun), and the meaning of *'esta'* (*'this'*, determiner) from *'está'* (*'is'*, verb) by marking the stressed syllable in the last word. As we can see, the regular tagger fails on these words, but the expanded one is able to tag them satisfactorily.

	No	**he**	**tenido**	**tiempo**	**de**	**escribir**	**sobre**	**el**	**y**
	not	have	had	time	of	to write	about	it	and
EXPANDED	ADV	VERB	VERB	NOUN	ADP	VERB	ADP	PRON	CONJ
REGULAR	ADV	VERB	VERB	NOUN	ADP	VERB	ADP	DET	CONJ
	ya	**esta**	**estropeado**						
	already	is	broken						
EXPANDED	ADV	VERB	VERB						
REGULAR	ADV	DET	VERB						

□

To evaluate PoS taggers, accuracy (number of correctly predicted tags divided by the total number of tags in the test set) is commonly the most used metric. The accuracy obtained on the unknown words in the test set (words that did not occur in the training set) is also reported by some systems (e.g. Toutanova and Manning (2000)). As the state-of-the-art performance of taggers for language such as English is above 97% and improvements can be measured in terms of tenths or even hundredths of percentage points, reporting the error reduction percentage is also common.

Tools to build PoS taggers

There are a number of available implementations to train PoS taggers. To name a few available systems: maximum-entropy (Toutanova and Manning, 2000), rule-based (Brill, 1992), statistical (Brants, 2000) or neural network taggers (Plank, Søgaard, and Goldberg, 2016; Schmid, 1994). In this book we are mainly relying on the Brill (Chapters 4 and 6) and Toutanova and Manning, 2000 taggers (Chapters 5 and 7).

2.3 DEPENDENCY PARSING

Dependency parsing is the process of obtaining a representation of the syntactic structure of a sentence, consisting of a set of oriented

binary relations between words. Each dependency has a label which denotes the existing syntactic relation between a pair of words. A formal definition of dependency parse can be consulted in Definition 2.

DEFINITION 2. A **dependency tree** for a sentence w is an edge-labeled directed tree $T = (V, E)$ where $V = \{0, 1, 2, \ldots, n\}$ is the set of nodes and $E = V \times D \times V$ is the set of labeled arcs. Each arc, of the form (i, d, j), corresponds to a syntactic **dependency** between the words w_i and w_j; where i is the index of the **head** word, j is the index of the **child** word and d is the **dependency type** representing the kind of syntactic relation between them. Following standard practice, we use node 0 as a dummy root node that acts as the head of the syntactic root(s) of the sentence.

□

Parsing is another core problem in natural language processing, with many potential applications given its ability to obtain the syntactic structure of a piece of text. Traditionally, the NLP community addressed this challenge mainly from a constituent parsing approach (Collins, 1997; Miller, 1995; Titov and Henderson, 2007), but in the last decade dependency parsing has gained increased interest (Andor et al., 2016; Chen and Manning, 2014; Dyer et al., 2015; Gómez-Rodríguez, Sartorio, and Satta, 2014; McDonald et al., 2013; McDonald et al., 2005; Nivre, 2003; Nivre et al., 2007). In the particular case of sentiment analysis, representing texts as dependency trees is more convenient, since it is possible to capture dependency relations like: *'to which word is affecting this adjective'*, *'what is the scope of this negation'* or *'which two sentences is relating this adversative subordinate clause?'*. Example 9 draws a valid dependency tree for our running example, where we can see that *'very'* is intensifying the adjective *'handsome'*. *'Very handsome'* is being negated and *'he has something that I really like'* is an adversative subordinate clause.

EXAMPLE 9 (A dependency tree). Figure 1 shows a valid dependency tree for our running example.

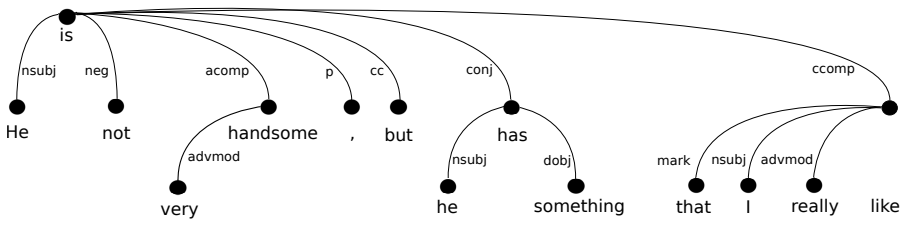


Figure 1: Example of a valid dependency tree for the running example: *'He is not very handsome, but he has something that I really like'*, following the McDonald et al. (2013) guidelines. For simplicity, we omit the dummy root in the figures.

□

We will write $i \xrightarrow{d} j$ as shorthand for $(i, d, j) \in E$ and we will omit the dependency types when they are not relevant.

The two most used metrics to measure the performance of parsers are:

- *Labeled Attachment Score* (LAS): Percentage of words that have their head and their dependency type correctly assigned.
- *Unlabeled Attachment Score* (UAS): Percentage of words that have their head correctly assigned.

CoNLL-X representation

The CONLL format (Buchholz and Marsi, 2006) for dependency parsing is a standard tabular representation to create the corpora used to train dependency parsers. Each token occupies a line and the content of the columns is as follows:

1. ID: The position in the sentence. The ID=0 is reserved for the dummy root.
2. FORM: The token itself.
3. LEMMA: The canonical word of the token.
4. CPOSTAG: The coarse PoS tag.
5. POSTAG: The fine-grained PoS tag.
6. FEATS: Additional morphological and syntactic information.
7. HEAD: The head of the token.
8. DEPREL: The dependency type between the token and its head.

Example 10 illustrates the graph showed in Example 9 in CONLL -2006 format.

EXAMPLE 10 (A dependency tree in CONLL format). A tabular output for the dependency tree drawn in Figure 1 according to the CONLL format.

id	form	lemma	cpostag	postag	feats	head	deprel
1	He	he	pronoun	_	_	2	nsubj
2	is	be	verb	_	_	0	root
3	not	not	adv	_	_	2	neg
4	very	very	adv	_	_	5	advmod
5	handsome	handsome	adv	_	_	2	acomp
6	,	,	punct	_	_	2	p
7	but	but	conj	_	_	2	cc
8	he	he	pronoun	_	_	9	nsubj
9	has	have	verb	_	_	2	conj
10	something	something	adv	_	_	9	dobj
11	that	that	sconj	_	_	14	mark
12	I	I	pronoun	_	_	14	nsubj
13	really	really	adv	_	_	14	advmod
14	like	like	verb	_	_	2	ccomp

□

Tools to build parsers

There are a number of tools to build data-driven dependency parsers (Andor et al., 2016; Chen and Manning, 2014; Martins, Almeida, and Smith, 2013; Rasooli and Tetreault, 2015). In this dissertation we are relying on MaltParser (Nivre et al., 2007), a system for automatically training dependency parsers that includes most of the traditional algorithms, such as the ones by Covington (2001), Nivre (2008b) or Gómez-Rodríguez and Nivre (2010). To optimize the model, we are relying on MaltOptimizer (Ballesteros and Nivre, 2012), a freely available tool developed to facilitate parser optimization with Maltparser.

There are many treebanks (Beek et al., 2002; Böhmová et al., 2003; Buchholz and Marsi, 2006; Džeroski et al., 2006; Zeman et al., 2012) that can be used to train such parsers depending on the target language, but in this dissertation we will be relying on three different collections: (1) Ancora (Taulé, Martí, and Recasens, 2008) a treebank that includes training and test data for Spanish and Catalan languages, (2) Universal Dependency Treebanks (UT) v2.0 (McDonald et al., 2013) a universal treebank for up to 10 languages that includes training, development and test sets and (3) Universal Dependencies (UD) v1.3 (Nivre et al., 2016), a revised version of the McDonald et al. (2013) treebanks, that contains revised guidelines and includes many more languages.

2.4 MULTILINGUAL PARSING

Multilingual natural language processing relying on parsing usually required to train a different parser for every target language, which complicates scalability and the maintenance of the multilingual pipeline. In this section we briefly describe a novel technique to train bilingual and multilingual parsers.

The goal of training multilingual parsers consists in creating a single model capable of parsing many languages, without the need of using any language identification tool. The basic idea to keep in mind is that a bilingual or multilingual parser should keep a robust performance on every language, in the target test set, in comparison to the corresponding monolingual parsers. To illustrate this, we use the Universal Dependency Treebanks v2.0 (McDonald et al., 2013), a set of CONLL -format treebanks for ten languages, annotated with common criteria. They include two versions of PoS tags: universal tags (Petrov, Das, and McDonald, 2011) in the CPOSTAG column, and a refined annotation with treebank-specific information in the POSTAG column. Some of the latter tags are not part of the core universal set, and they can denote linguistic phenomena that are language-specific, or simply phenomena that not all the corpora have annotated in the same way. We briefly describe the process and results below these lines. The same procedure is applicable to create multilingual taggers, and an example is briefly introduced in §2.4.1.

To train monolingual parsers (our baseline to measure the quality of the bilingual and multilingual parsers), we used the official training and dev-set splits provided with the corpora. For the bilingual models, for each pair of languages L_1, L_2; we simply merged their training sets into a single file acting as a training set for $L_1 \cup L_2$, and we did the same for the development sets. The test sets were not merged because comparing the bilingual parsers to monolingual ones requires evaluating each bilingual parser on the two corresponding monolingual test sets.

To build the models, we relied on MaltParser optimizing the models with MaltOptimizer. This system works in three phases: *Phase 1* and *2* choose a parsing algorithm by analyzing the training set, and performing experiments with default features. *Phase 3* tunes the feature model and algorithm parameters. We hypothesize that the bilingual models will learn a set of features that fits both languages, and check this hypothesis by evaluating on the test sets. We propose two training configurations:

1. a *treebank-dependent tags* configuration where we include the information in the POSTAG column.
2. a *universal tags only* configuration, where we do not use this information, relying only on the CPOSTAG column.

Information that could be present in FEATS or LEMMA columns is not used in any case. This methodology plans to answer two research questions that are relevant to address multilinguality in sentiment analysis:

1. can we train bilingual parsers with good accuracy by merging harmonized training sets?
2. is it essential that the tagsets for both languages are the same, or can we still get accuracy gains from fine-grained PoS tags (as in the monolingual case) even if some of them are treebank-specific?

To ensure a fair comparison between monolingual and bilingual models, we chose to optimize the models from scratch with MaltOptimizer, expecting it to choose the parsing algorithm and feature model which is most likely to obtain good results. We observed that the selection of a bilingual parsing algorithm was not necessarily related with the algorithms selected for the monolingual models. The system sometimes chose an algorithm for a bilingual model that was not selected for any of the corresponding monolingual models.

In view of this, and as it is known that different parsing algorithms can be more or less competitive depending on the language (Nivre, 2008a), we ran a control experiment to evaluate the models setting the same parsing algorithm for all cases, executing only *phase 3* of MaltOptimizer. We chose the arc-eager parser for this experiment, as it was the algorithm that MaltOptimizer chose most frequently for the monolingual models in the previous configuration. The aim was to compare the accuracy of the bilingual models with respect to the monolingual ones, when there is no variation on the parsing algorithm between them. The results of this control experiment were very similar to those of the original experiment.

Evaluating bilingual parsers

Table 1 compares the accuracy of bilingual models to that of monolingual ones, under the *treebank-dependent tags* configuration. Each table cell shows the accuracy of a model, in terms of LAS and UAS. Cells in the diagonal correspond to monolingual models (the baseline), with the cell located at row i and column i representing the result obtained by training a monolingual parser on the training set of language L_i, and evaluating it on the test set of the same language L_i. Each cell outside the diagonal (at row i and column j, with $j \neq i$) shows the results of training a bilingual model on the training set for $L_i \cup L_j$, evaluated on the test set of L_i.

As we can see, in a large majority of cases, bilingual parsers learn to parse two languages with no statistically significant accuracy loss with respect to the corresponding monolingual parsers ($p < 0.05$ with Bikel's randomized parsing evaluation comparator). This happened in 74 out of 90 cases when measuring UAS, or 69 out of 90 in terms

	de	*en*	*es*	*fr*	*id*	*it*	*ja*	*ko*	*pt-br*	*sv*
de	78.27 84.03	78.01^{-} 84.08^{+}	77.82^{-} 83.82^{-}	77.83^{-} 83.55^{--}	77.84^{-} 83.85^{-}	78.10^{-} 84.12^{+}	77.86^{-} 83.88^{-}	77.94^{-} 83.63^{-}	78.13^{-} 83.87^{-}	78.60^{+} 84.38^{+}
en	89.37^{+} 91.02^{+}	89.36 91.02	89.46^{+} 91.09^{+}	89.38^{+} 91.06^{+}	89.69^{++} 91.32^{++}	89.82^{++} 91.47^{++}	89.43^{+} 91.10^{+}	89.63^{++} 91.32^{++}	89.60^{++} 91.24^{+}	89.11^{--} 90.79^{--}
es	80.85^{+} 85.17^{+}	81.08^{++} 85.27^{++}	80.60 84.75	80.95^{+} 85.15^{+}	81.16^{+} 85.00^{+}	80.92^{+} 85.13^{+}	81.41^{++} 85.52^{++}	81.49^{++} 85.39^{++}	79.96^{-} 84.70^{-}	81.26^{++} 85.42^{++}
fr	79.01^{-} 84.17^{-}	79.39^{+} 84.49^{+}	79.36^{+} 84.56^{+}	79.29 84.47	79.61^{+} 84.32^{-}	79.34^{+} 84.41^{-}	79.16^{-} 84.34^{-}	79.36^{+} 84.72^{+}	79.09^{-} 83.98^{-}	79.66^{+} 84.84^{+}
id	75.72^{--} 81.73^{--}	77.19^{-} 82.66^{--}	77.12^{-} 82.72^{-}	77.15^{-} 82.66^{-}	77.69 83.38	78.29^{+} 84.09^{+}	77.60^{-} 83.18^{-}	76.68^{--} 82.16^{--}	77.45^{-} 82.96^{-}	77.01^{--} 82.59^{--}
it	82.62^{--} 86.14^{--}	83.17^{--} 86.46^{--}	83.12^{--} 86.78^{-}	83.10^{--} 86.69^{-}	83.74^{--} 86.73^{--}	84.40 87.54	84.62^{+} 87.48^{-}	84.79^{+} 87.46^{-}	83.70^{-} 87.39^{-}	84.55^{+} 87.23^{-}
ja	76.53^{--} 83.77^{-}	76.24^{--} 83.89^{-}	76.61^{--} 84.26^{-}	76.32^{--} 84.05^{--}	75.18^{--} 83.08^{--}	77.05^{-} 83.97^{-}	77.46 84.34	76.89^{-} 83.65^{-}	76.69^{-} 83.97^{-}	76.89^{-} 84.17^{-}
ko	86.13^{--} 90.61^{--}	88.30^{+} 92.16^{+}	87.91^{+} 92.00^{-}	88.49^{+} 92.35^{+}	85.86^{--} 90.19^{--}	88.72^{++} 92.55^{+}	87.14^{--} 91.89^{-}	87.83 92.12	86.75^{--} 91.39^{--}	88.68^{-} 92.39^{-}
pt-br	84.83^{-} 87.18^{-}	85.06^{+} 87.19^{-}	84.99^{+} 87.27^{+}	84.97^{+} 87.17^{-}	85.10^{+} 87.35^{-}	85.43^{++} 87.68^{++}	84.95^{+} 87.13^{-}	85.12^{+} 87.35^{-}	84.88 87.39	85.25^{++} 87.43^{++}
sv	81.71^{--} 86.01^{--}	82.01^{--} 86.39^{-}	82.03^{-} 86.55^{-}	81.92^{--} 86.28^{--}	82.34^{-} 86.69^{-}	82.63^{+} 86.55^{-}	82.81^{+} 86.92^{+}	82.94^{++} 86.83^{-}	82.19^{-} 86.39^{-}	82.48 86.92

Table 1: LAS/UAS performance on the Universal Dependency Treebanks test sets by the mono and bilingual parsers. For each cell, its (row,column) pair indicates the language(s) with which the model was trained, with the row corresponding to the language where it was evaluated. '++' and '+' indicate that the improvement in performance obtained by the bilingual model is statistically significant or not, respectively. '--' and '-' correspond to significant and not significant *decreases* in accuracy.

of LAS. Therefore, in most cases where we are applying a parser to texts in a given language, adding a second language comes for free in terms of accuracy.

More strikingly, there are many cases where bilingual parsers outperform monolingual ones, even in this evaluation on purely monolingual datasets. In particular, there are 12 cases where a bilingual parser obtains statistically significant gains in LAS over the monolingual baseline, and 9 cases with significant gains in UAS. This clearly surpasses the amount of significant gains to be expected by chance, and applying the Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995) to correct for multiple comparisons with a maximum false discovery rate of 20% yields 8 significant improvements in LAS and UAS. Therefore, it is clear that there is synergy between datasets: in some cases, adding annotated data in a different language to our training set can actually improve the accuracy that we obtain in the *original* language. This opens up interesting research potential in using confidence criteria to select the data that can help parsing in this way, akin to what is done in self- training approaches (Chen, Wu, and Isahara, 2008; Goutam and Ambati, 2011).

Comparing the results by language, we note that the accuracy on the English and Spanish datasets almost always improves when adding a second treebank for training. Other languages that tend to get improvements in this way are French and Portuguese. There seems to be a rough trend towards the languages with the largest training corpora benefiting from adding a second language, and those with the smallest corpora (e.g. Indonesian, Italian or Japanese) suffering accuracy loss, likely because the training gets biased towards the second language.

Training bilingual models containing a significant number of non-overlapping treebank-dependent tags tends to have a positive effect. English and Spanish are two of the clearest examples of this. As shown in Table 2, which shows a complete report of shared PoS tags for each pair of languages under the *treebank-dependent* tags configuration, English only shares 1 PoS tag with the rest of the corpora under the said configuration, except for Swedish, with up to 5 tags in common; and the *en-sv* model is the only one suffering a significant loss on the English test set. Similar behavior is observed on Spanish: *sv* (0), *en* (1), *ja* (10) and *ko* (12) are the four languages with fewest shared PoS tags, and those are the four that obtained a significant improvement on the Spanish evaluation; while with *pt-br*, with 15 shared PoS tags, we lose accuracy. The validity of this hypothesis is reinforced by an experiment where we differentiate the universal tags by language by appending a language code to them (e.g. EN_NOUN for an English noun). An overall improvement was observed with respect to the bilingual parsers with non-disjoint sets of features.

	de	*en*	*es*	*fr*	*id*	*it*	*ja*	*ko*	*pt-br*	*sv*
de	16	1	14	14	14	13	10	12	14	0
en		45	1	1	1	1	1	1	1	5
es			24	14	14	13	10	12	15	0
fr				14	14	13	10	12	14	0
id					14	13	10	12	14	0
it						13	10	12	13	0
ja							763	10	10	0
ko								20	12	0
pt-br									15	0
sv										25

Table 2: Shared language-specific tags between pairs of languages on the Universal Dependency Treebanks v2.0

While all these experiments have been performed on sentences with gold PoS tags, preliminary experiments assuming predicted tags instead show analogous results: the absolute values of LAS and UAS are slightly smaller across the board, but the behavior in relative terms is the same, and the bilingual models that improved over the monolingual baseline in the gold experiments keep doing so under this setting.

On the other hand, Table 3 shows the performance of the monolingual and bilingual models under the *universal tags only* configuration. The bilingual parsers are also able to keep an acceptable accuracy with respect to the monolingual models, but significant losses are much more prevalent than under the *treebank-dependent tags* configuration.

Putting both tables together, our experiments clearly suggest that not only treebank-specific tags do not impair the training of bilingual models, but they are even beneficial, supporting the idea that using partially treebank-dependent tagsets helps multilingual parsing. We hypothesize that this may be because complementing the universal information at the syntactic level with language-specific information at the lower levels (lexical and morphological) may help the parser identify specific constructions of one language that would not benefit from the knowledge learned from the other, preventing it from trying to exploit spurious similarities between languages. This explanation is coherent with work on delexicalized parser transfer (Lynn et al., 2014) showing that better results can be obtained using disparate languages than closely-related languages, as long as they have common syntactic constructions. Thus, using universal PoS tags to train multilingual parsers can be, surprisingly, counterproductive.

	de	*en*	*es*	*fr*	*id*	*it*	*ja*	*ko*	*pt-br*	*sv*
de	74.07	72.04^{--}	74.51^{+}	74.44^{+}	73.68^{-}	73.76^{-}	73.90^{-}	74.30^{+}	74.29^{+}	74.76^{++}
	79.77	77.52^{--}	79.95^{+}	79.83^{+}	79.24^{-}	79.44^{-}	79.83^{+}	79.76^{-}	79.71^{-}	80.25^{+}
en	88.46^{+}	88.35	88.65^{++}	88.39^{+}	88.61^{++}	88.68^{++}	88.65^{++}	88.61^{++}	88.65^{++}	88.50^{+}
	90.35^{+}	90.27	90.54^{++}	90.26^{-}	90.47^{++}	90.53^{++}	90.49^{++}	90.43^{++}	90.55^{++}	90.43^{++}
es	79.66^{--}	78.78^{--}	80.54	79.59^{--}	78.98^{--}	79.84^{--}	79.59^{--}	79.80^{--}	79.74^{--}	79.09^{--}
	83.81^{--}	82.94^{--}	84.35	83.26^{--}	82.79^{--}	83.79^{--}	83.53^{--}	83.57^{-}	83.76^{--}	83.28^{--}
fr	78.43^{+}	78.10^{-}	78.63^{+}	78.40	77.79^{-}	78.60^{+}	79.11^{+}	78.22^{-}	78.56^{+}	78.83^{+}
	83.26^{-}	82.77^{-}	83.38^{-}	83.40	82.85^{-}	83.50^{+}	84.03^{+}	83.05^{-}	83.45^{+}	83.73^{+}
id	74.46^{--}	74.65^{--}	77.09^{--}	76.23^{--}	78.31	77.86^{-}	77.10^{--}	75.58^{--}	76.90^{--}	78.34^{+}
	80.87^{--}	80.21^{--}	82.81^{--}	81.78^{--}	83.81	83.52^{-}	82.68^{--}	81.20^{--}	82.50^{--}	83.83^{+}
it	82.27^{--}	82.13^{--}	82.24^{--}	82.75^{--}	82.65^{--}	83.88	83.04^{--}	83.77^{-}	83.07^{--}	83.47^{-}
	85.40^{--}	85.38^{--}	85.36^{--}	86.31^{--}	85.45^{--}	86.68	85.83^{--}	86.30^{-}	86.21^{--}	86.33^{-}
ja	69.41^{--}	68.88^{--}	69.28^{--}	69.24^{--}	69.73^{--}	70.22^{--}	70.87	69.73^{--}	69.24^{--}	70.02^{-}
	79.62^{--}	79.21^{--}	79.45^{--}	80.11^{--}	79.58^{--}	79.58^{--}	81.16	80.23^{-}	79.37^{--}	80.47^{--}
ko	84.40^{--}	84.82^{--}	85.40^{--}	84.59^{--}	84.74^{--}	86.79^{-}	86.21^{--}	87.52	86.29^{--}	86.40^{--}
	89.61^{--}	90.00^{--}	90.77^{--}	89.88^{--}	90.00^{--}	91.39^{-}	91.46^{--}	92.00	90.92^{--}	91.19^{--}
pt-br	83.40^{-}	82.76^{--}	83.56^{-}	83.72^{-}	83.08^{--}	83.95^{+}	83.80^{-}	84.16++	83.83	84.28++
	85.78^{-}	85.01^{--}	85.82^{-}	85.85^{-}	85.38^{--}	86.15^{+}	85.93^{-}	86.33^{+}	86.11	86.41++
sv	79.65^{--}	79.61^{--}	79.75^{--}	80.46^{-}	80.94^{+}	81.06^{+}	81.19^{+}	81.11^{+}	80.89^{-}	80.93
	84.14^{--}	84.42^{--}	84.46^{--}	84.88^{-}	85.14^{-}	85.51^{+}	85.29^{-}	85.14^{-}	85.05^{-}	85.32

Table 3: Performance on the Universal Dependency Treebanks v2.0 test sets using the gold CPOSTAG information

Adding more languages

To show that our approach works when more languages are added, we created a quadrilingual parser using the romanic languages and the fine PoS tag set. The results (LAS/UAS) on the monolingual sets were: 80.18/84.64 (*es*), 79.11/84.29 (*fr*), 82.16/86.15 (*it*) and 84.45/86.80 (*pt*). In all cases, the performance is almost equivalent to the monolingual parser. Ammar et al. (2016) has shown that this idea can be also adapted to universal parsing.

2.4.1 *Code-switching parsing*

Our bilingual parsers also show robustness on texts exhibiting code-switching. Unfortunately, there are no syntactically annotated code-switching corpora, so we could not perform a formal evaluation. We did perform informal tests, by running the Spanish-English bilingual parsers on some such sentences. We observed that they were able to parse the English and Spanish parts of the sentences much better than monolingual models. This required training a bilingual tagger, which we did with the free distribution of the Stanford tagger (Toutanova and Manning, 2000); merging the Spanish and English corpora to train a combined bilingual tagger. Under the *universal tags only* configuration, the multilingual tagger obtained 98.00% and 95.88% over

the monolingual test sets. Using treebank-dependent tags instead, it obtained 97.19% and 93.88% over the monolingual test sets. Example 11 shows the output of the bilingual tagger on a code-switching sentence, compared against the corresponding monolingual English and Spanish taggers. Example 12 illustrates the resulting dependency trees for the same sentence, when different combinations of monolingual and bilingual taggers and parsers are combined.

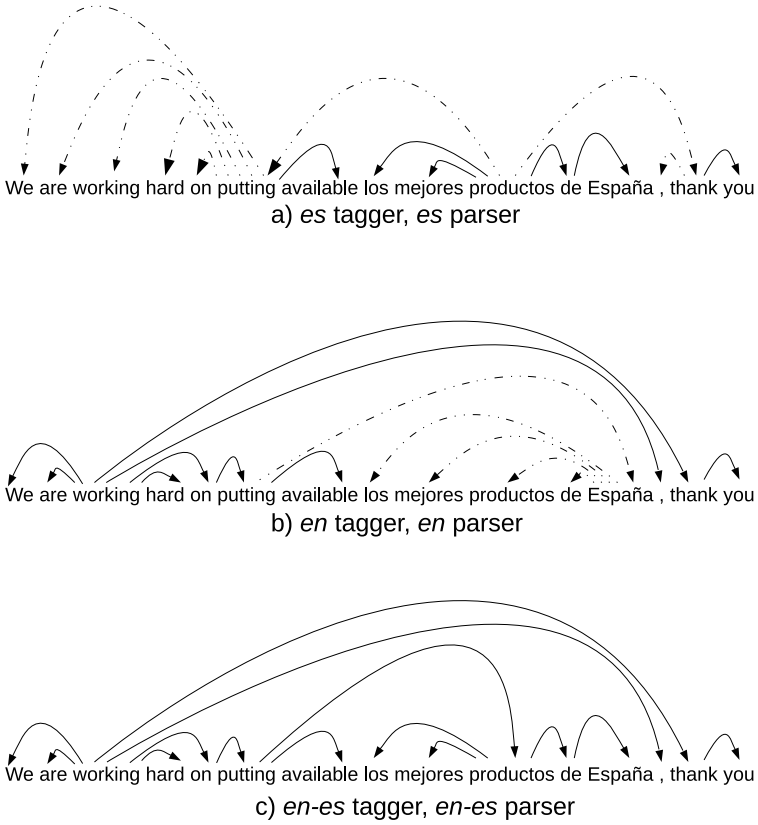
EXAMPLE 11. Performance of the bilingual English-Spanish (*en-es*) tagger with respect to a monolingual English (*en*) and Spanish (*es*) tagger.

Tagger	**We**	**are**	**working**	**hard**	**on**	**putting**	**available**	**los**	**mejores**
es	NOUN	NOUN	NOUN	NOUN	NOUN	NOUN	ADJ	DET	ADJ
en	PRON	VERB	VERB	ADV	ADP	VERB	ADJ	X	X
es-en	PRON	VERB	VERB	ADV	ADP	VERB	ADJ	DET	ADJ
	productos	**de**	**España**	**,**	**thank**	**you**			
es	NOUN	ADP	NOUN	.	X	X			
en	X	X	NOUN	.	VERB	PRON			
es-en	NOUN	ADP	NOUN	.	VERB	PRON			

The example uses the universal tagset proposed by Petrov, Das, and McDonald (2011).

□

EXAMPLE 12 (Dependency parsing on an English-Spanish code-switching sentence by different models). We illustrate how using bilingual parsers (and taggers) affects the accuracy for the code-switching sentence: *'We are working hard on putting available los mejores productos de España, thank you'* (*'We are working hard on putting available the best products of Spain, thank you'*).



En-es refers to a bilingual tagger or parser, and *en* and *es* to a monolingual English or Spanish model. We illustrate the output when we make different combinations of those. Dotted/dashed lines represent incorrectly-parsed dependencies.

□

Tagger	Parser	las	uas
en	*en*	37.82	44.23
es	*es*	27.56	41.03
en-es	*en*	66.03	78.85
en-es	*es*	67.95	77.56
en-es	*en-es*	**87.18**	**92.31**

Table 4: LAS/UAS performance on a code-switching Universal Dependency treebank composed of 10 sentences

Finally, Table 4 shows the performance on a tiny code-switching treebank built on top of ten normalized tweets.[4] This confirms that

4 The code-switching treebank follows the Universal Treebank v2.0 annotations. It can be obtained by emailing david.vilares@udc.es

monolingual pipelines perform poorly. Using a bilingual tagger helps improve the performance, thanks to accurate tags for both languages, but a bilingual parser is needed to push both LAS and UAS up to state-of-the-art levels.

2.5 EVALUATION METRICS FOR SENTIMENT ANALYSIS

To evaluate polarity classification systems researchers usually rely on precision (P) (Equation 1), recall (R) (Equation 2), f1-measure (F1) (Equation 3) and accuracy (Equation 4) when dealing with systems that provide a discrete output:

$$Precision_i = \frac{tp_i}{tp_i + fp_i} \tag{1}$$

$$Recall_i = \frac{tp_i}{tp_i + fn_i} \tag{2}$$

$$F1_i = \frac{2 \cdot Precision_i \cdot Recall_i}{Precision_i \cdot Recall_i} \tag{3}$$

$$Accuracy = \frac{\sum_{i=0}^{n} tp_i + \sum_{i=0}^{n} tn_i}{\sum_{i=0}^{n} tp_i + \sum_{i=0}^{n} tn_i + \sum_{i=0}^{n} fp_i + \sum_{i=0}^{n} fn_i} \tag{4}$$

where:

- tp_i is the true positive classifications for class i.
- fp_i is the false positive classifications for class i.
- tn_i is the true negative classifications for class i.
- fn_i is the false negatvie classifications for class i.
- n is the total number of classes.

The first three are usually computed for each category i to be analyzed, meanwhile accuracy is commonly used to give global results. To provide global results for precision, recall and f1-measure the option followed in this thesis is to compute either the micro- or macro-average.

For systems such as Sentistrength, which provide a numerical output, the *de facto* standard to measure the quality of the systems has been: *pearson correlation* (Equation 5), *accuracy* and also a relaxed accuracy, *+/-1 correct classification* (Equation 6).

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} = \frac{E[(X - \mu_x)(Y - \mu_y)]}{\sigma_x \sigma_y} \tag{5}$$

where:

- μ_{xy} is the covariance of (X, Y).

- μ_x is the typical desviation of variable X.
- μ_y is the typical variation of variable Y.
- E is the expectation.

$$+/1 \text{ correct} = \frac{\sum_{i=0}^{n} rtp_i}{\sum_{i=0}^{n} tp_i + \sum_{i=0}^{n} tn_i + \sum_{i=0}^{n} fp_i + \sum_{i=0}^{n} fn_i} \quad (6)$$

where:

- rtp_i is the number fo true positive for the class i, asumming that the classes are a discrete and ordered set and a correct classification for an instance in an specific class also includes to be assigned into any of the collateral categories.

Part II

KNOWLEDGE-BASED COMPOSITIONAL STRATEGIES

3

A LEXICAL, UNSUPERVISED, KNOWLEDGE-BASED APPROACH

In this chapter we adapt SentiStrength (Thelwall, Buckley, and Paltoglou, 2012) to the Spanish language. SentiStrength is a lexicon-based approach for sentiment analysis that handles relevant linguistic phenomena for this task (e.g. negation or intensification), by using window-based rules over texts. This chapter describes how we adapted an existing baseline of the system for Spanish, how we built a training and evaluation corpus and how we improved the old model's performance. In Chapter 9 we will be using it to analyze in real time tweets referring to the main Spanish politicians.

The new version of Spanish SentiStrength is available for research[1] and it can be tried online[2].

3.1 DESCRIPTION

SentiStrength was created to partly fill the gap of many traditional tools that were able to extract the polarity but not the sentiment strength from short informal texts in English (Thelwall, Buckley, and Paltoglou, 2012; Thelwall et al., 2010). Whilst sometimes extracting the polarity might be enough, in some situations Thelwall et al. (2010) point out that extracting the strength of the opinions presents advantages for basic research on the role of emotion in online communication (Derks, Fischer, and Bos, 2008) or discussions (Balahur, Kozareva, and Montoyo, 2009). In this context, SentiStrength was one of the first fine-grained sentiment analysis systems. Among its advantages, it is able to take into account the grammar style present in texts for the purpose at hand (e. g. repetition of characters in a word might mean a way of explicit intensification) and provides a dual score output which gives an idea of the human behavior. In particular, it follows a dual score approach where each text receives a score from 1 (no positivity) to 5 (very strong positivity) and a second score from -1 (no negativity) to -5 (very strong negativity). We will be referring to these scores according to the notation $(positive, negative)$ or the *italic* name *sentistrength*.

SentiStrength works as a window-based model that uses rules to deal with some of the most relevant phenomena in sentiment analysis. It follows a maximum sentiment approach instead of summing up individual so's, i. e. given two phrases with the *sentistrength* $(2, -2)$ and

1 http://sentistrength.wlv.ac.uk/SpanishSentiDataDavidVilares.zip
2 http://sentistrength.wlv.ac.uk/#Non-English

$(2, -3)$ the global *sentistrength* is $(2, -3)$ and not $(4, -5)$. A detailed description for all of SentiStrength capabilities and how they work can be found in Thelwall, Buckley, and Paltoglou (2012) and Thelwall et al. (2010).

3.1.1 *Sentiment dictionary*

At the core of lexical sentiment analysis algorithms there are lists of sentiment-related terms. Although there are several Spanish sentiment lexicons, they mainly distinguish between positive and negative words (Martínez Cámara et al., 2014; Saralegi and San Vicente, 2013) and do not use the same sentiment scale as SentiStrength. A first version of the SentiStrength Spanish dictionary (García and Thelwall, 2013) contained 1 409 subjective terms, each annotated with a sentiment strength of 1 to 5 or -1 to -5. These terms were mainly derived from LIWC (Pennebaker, Francis, and Booth, 2001), which contains a psychological text analysis resource with a Spanish variant. Additional terms were also added by SentiStrength's commercial users and from Spanish translations of other English resources (Bradley and Lang, 1999; Redondo et al., 2007).

We have extended this sentiment dictionary with a new dictionary of subjective adjectives, nouns, verbs and adverbs. Brooke, Tofiloski, and Taboada (2009) was considered as primary source for new terms to add. The semantic orientation of these terms range from 1 to 5 (if it is a positive term) and from -1 to -5 (if it is a negative one).

SentiStrength does not use typical natural language processing steps such as lemmatization, part-of-speech tagging or dependency parsing, because of its focus on short informal texts, such as tweets, in which non-standard spelling and grammar probably occur in the majority of texts. This is an issue for Spanish, as most nouns and adjectives vary with gender and number, making the direct matching of dictionary words to text words difficult. As a result, the Spanish dictionary might need to include four versions of most nouns to cope with their gender and number variations. Verb inflections differ even more than nouns. To deal with this issue, the dictionaries were expanded to include the common word form variants in each case. In particular, Spanish verbs are classified in three groups depending on the ending of its infinitive: *'ar'*, *'er'* or *'ir'*, and we inflected the verb forms with all the possible variations coming from Spanish verb tenses for such three groups (irregular verbs are not taken into account in this study), except irrealis conditional and subjective forms, since it has been reported that they need to be processed in a way the current implementation of SentiStrength cannot handle (Taboada et al., 2011; Trnavac and Taboada, 2012).

The improved version of SentiStrength with the new dictionaries was applied to the training set and discrepancies in the results ana-

lyzed to identify additional modifications. This resulted in the addition of new terms as well as the removal of wrong terms, such as homonyms with sentiment and non-sentiment meanings, and modifications of some of the original term strengths. The final Spanish lexicon was almost 20 times larger than the first one, with 26 752 entries (5 728 word forms plus verb, noun and adjective inflections).

3.1.2 *Additional sentiment files*

As previously mentioned, the Sentiment dictionary is the core of SentiStrength, as happens with other lexicon-based methods, but it also uses other lexicons with relevant content for the purpose at hand, typically referred to as additional *sentidata*.

- *Emoticons*. SentiStrength considered a list of 115 traditional emoticons to help detect sentiment (see also: Tang et al. (2014)). We have added 71 new emoticons from the training set.
- *Idioms*. Only 3 stock Spanish phrases that have a different sentiment than the individual words (e.g., *'Que tal'* (*'How are you'*) scores +2) were available. We have expanded it with 306 extra expressions, many involving verbs and their inflections.
- *Slang words*. Although the English version of SentiStrength employs a slang conversion table, the Spanish version did not. Therefore, a list of common Spanish slang and abbreviations has been created from the training set, using as vocabulary the words occurring at the Ancora corpus (Taulé, Martí, and Recasens, 2008). Many of the abbreviations had subjective connotations, such as *'tkm'* (*'te quiero mucho'* - *'I love you'*) and *'bs'* (*'beso'* - *'kiss'*). Some frequent Anglicisms were also translated into Spanish (e.g., *'VIP'* - *'persona muy importante'*).
- *Vocabulary*. This collection of words that are part of the language is used within SentiStrength's spelling correction algorithm. It was expanded with the words occurring at Ancora.
- *Booster words*. Only ten booster words to amplify or diminish the sentiments of subjective words were available. We expanded it using another booster collection (Brooke, Tofiloski, and Taboada, 2009) and new terms from the development set, mainly from South American Spanish (e. g. *'re'* and *'so'*), giving 169 terms.
- *Question words*. Spanish SentiStrength used five words (e. g. *'qué'*) to identify questions, which have modified sentiment rules. This list was extended to 20 terms. Acute accents are important for these because their absence turns many question words into conjunctions (e. g. *'cómo de bien lo hizo'* that is translated to English as *'how well we did it'*, could be understood as *'como de bien lo hizo'* if accents are omitted, which could translate to *'he did it really well'*), but they are usually omitted in Twitter, so the unaccented word forms were also included.

3.2 EXPERIMENTS

3.2.1 *Dataset*

A human-annotated corpus is needed to evaluate SentiStrength, with each text given a positive and a negative sentiment strength score. No published formal evaluations were available for Spanish, nor any human-coded Spanish corpora with positive and negative sentiment strength scores. In response, we created both a development and a test set from a large collection of Spanish tweets downloaded from the Twitter Streaming API in September 2014:

- *Development set*. A collection of 1 600 tweets was labeled by an expert annotator. The corpus was used to explore ways of improving the performance of SentiStrength, as explained in the following section.
- *Test set*. A collection of 1 600 tweets was manually labeled by Spanish computational linguists. To identify reliable coders, we first asked seven people to annotate a common set of 160 tweets. We then selected the three annotators that coded most consistently against each other, with Krippendorf's alpha coefficient of inter-coder consistency varying from 0.630 to 0.701 for negative sentiment and 0.625 to 0.726 for positive sentiment. These three annotators were then asked to independently label other 1 600 tweets for the test set, obtaining consistency from 0.486 to 0.660 for negative sentiment and 0.503 to 0.678 for positive sentiment. Three different strategies were used to combine the scores: the average; the maximum (assuming that annotators tend to be conservative); and the average after removing the minimum.

The corpus has been made available for research purposes and can be downloaded from this link.[3]

3.2.2 *Evaluation*

SentiStrength was evaluated by comparing its output to the results of the human coders on the test set of tweets (i. e. the gold standard). The optimal metric for comparisons is the simple Pearson correlation because it reflects the closeness between the prediction and the true value in cases where the prediction is not perfect. For completeness and comparisons with other systems, we also report the percentage of scores that equal the gold standard (i. e. precision), the percentage of scores that differ by at most 1 from the gold standard (called +/-1), and the trinary accuracy. The trinary metric uses only three classes:

3 http://sentistrength.wlv.ac.uk/SpanishTweetsTestAndDevelopmentSetsDavidVilares.zip

positive (POS), neutral/mixed (NEU) (either no subjectivity or equal positive and negative scores) and negative (NEG). The new version of SentiStrength substantially outperforms the original version both for positive and negative sentiment in terms of the key correlation metric, performs moderately better on the trinary metric and performs slightly better overall for precision but slightly worse for +/-1 (Table 5). The reason for the moderately worse performance on negative sentiment +/-1 is probably because the old version of Spanish SentiStrength had a relatively small set of negative terms and mostly assigned the minimum no negative score. In fact the strategy of assigning -1 to all tweets would get a high score in the +/-1 metric (89.6% on the version of the test set obtained by averaging annotator scores average test set) due to the relative scarcity of negative tweets.

Measure	**Average**		**Maximum**		**Minimum**	
	New	**Old**	**New**	**Old**	**New**	**Old**
Positive correlation	**0.437**	0.304	**0.437**	0.326	**0.437**	0.294
Positive correct (%)	**51.4**	47.6	**47.3**	44.3	**51.3**	46.7
Positive+/-1 (%)	**79.9**	79.8	**78.1**	76.1	**79.0**	78.7
Negative correlation	**0.421**	0.351	**0.423**	**0.349**	**0.417**	0.341
Negative correct (%)	**63.4**	63.1	**52.4**	51.6	**63.6**	62.0
Negative+/-1 (%)	86.2	**89.0**	79.3	**80.1**	84.5	**86.6**
Trinary evaluation (%)	**54.5**	50.9	**52.5**	49.8	**55.2**	51.1

Table 5: Spanish SentiStrength performance (the old and new models) under the default setup on the three versions of test set of 1600 human-coded tweets.

SentiStrength includes a number of options to configure the behavior of the algorithm (Thelwall, Buckley, and Paltoglou, 2012), such as the maximum number of terms allowed between a negating word and a sentiment word. We used the development set to assess whether changes in these parameters could improve the overall SentiStrength results. We optimized the choice of parameters via a greedy search based upon performance (using the main correlation metric) on the development set. For example, the experiments on the development set indicated that flipping negated negative words to positive was better than neutralizing them.

Option disabled	Positive correlation	Positive correct	Positive correct +/-1	Negative correlation	Negative correct	Negative correct +/-1
Spelling correction	0.416	52.312	80.062	0.406	63.625	86.438
Questions reduce negative sentiment	0.436	51.437	79.812	0.409	**64.75**	**87.688**
Multiple letters boost sentiment	0.434	**52.375**	80.312	0.423	63.437	86.312
Booster list	0.433	51.750	**80.625**	0.415	63.462	86.625
Dictionary	0.427	51.812	79.875	0.410	63.562	86.375
Multiple positive words are boosted	0.437	51.562	80.125	0.421	63.375	86.188
Emoticon list	0.437	51.437	79.938	0.421	63.375	86.188
Negating positive neutralised emotion	0.435	50.500	79.500	**0.424**	63.937	86.188
Multiple negative words are boosted	0.437	51.437	79.938	0.421	63.375	86.25
Ignore booster words after negators	0.437	51.437	79.938	0.421	63.375	86.188
Exclamation marks count as +2	0.437	51.375	79.938	0.421	63.375	86.188
Negating negative neutralises emotion	**0.438**	51.500	79.938	0.417	62.375	85.625
Idiom list	**0.438**	51.500	79.875	0.416	63.250	86.188
None (default configuration)	0.437	51.437	79.938	0.421	63.375	86.188

Table 6: Spanish SentiStrength performance with a variety of options individually disabled on the test set (average scores).

Table 6 details the performance of the new SentiStrength on the test set with different running configurations, activating or deactivating one option at a time. For simplicity, the remaining results use only the average of the scores of the three annotators. Most variations in performance are small. Table 6 also shows that the best configuration on the development set (the one disactivating the option Negating negative neutralizes emotion) achieved the highest positive correlation and an acceptable performance on the rest of the metrics. This reinforces the competitiveness of this model for analyzing real texts.

3.3 CONCLUSION

This chapter extended the sentiment strength detection program SentiStrength for the Spanish language, taking as the starting point an existing baseline. We collected and expanded resources that feed the system and evaluated the performance of different phenomena present in web opinions, that potentially reflect some kind of emotion and that can be handled by SentiStrength. To do this, a Spanish Twitter corpus annotated according to the dual SentiStrength score was built. Experimental results show that the new Spanish version clearly improves over the existent baseline. We also evaluated different setups, that disabled the treatment of individual phenomena, but it was observed that in general the impact is small.

Sentistrength is a simple and robust option when we want to perform fast large-scale data analysis in real time consuming few resources (see Chapter 9). However, it is limited in the range of the phenomena that can be handled or even in the scope of the rules, that only consider very shallow structure. These problems will be tackled in Chapters 4 and 5.

4

A SYNTACTIC KNOWLEDGE-BASED APPROACH FOR MONOLINGUAL SENTIMENT ANALYSIS

In Chapter 3, we proposed a purely lexicon-based approach for fine-grained classification of Spanish short texts. However, as happens to other lexicon-based systems (Taboada et al., 2011), it cannot take into account the relations between words because it cannot interpret the syntactic structure of texts. To overcome these limitations, it is common to implement heuristics to simulate a comprehension of negation, intensification and other linguistic constructions, but these often fail, given the complexity of natural languages. As an alternative, in the present chapter we introduce a richer linguistic approach that obtains the syntactic structure of sentences by means of a dependency parser. This structure is then used to address three of the most significant linguistic constructions for determining the semantic orientation: intensification, subordinate adversative clauses and negation. We also introduce a semi-automatic domain adaptation method to improve the accuracy in specific application domains, one of the most important weaknessess of knowledge-based methods with respect to machine learning models. By enriching semantic dictionaries using machine learning methods to adapt the semantic orientation of their words to a particular field, we show how the proposed methods can achieve state-of-the-art results. We will be refering to the system presented in this chapter as Spanish Syntactic Sentiment Analysis (SSSA).

A model implementing this approach can be found as a part of the miopia library[1].

4.1 DESCRIPTION

Many SA systems do not take into account the relations between words because they cannot interpret the syntactic structure of texts. As an alternative, in this chapter we propose an unsupervised method for determining the semantic orientation of texts written in Spanish based on their dependency structure.

As a first step, texts are preprocessed according to §2.1. As a second step, we tokenize sentences and words to then apply PoS tagging. The next step consists in running the Brill (1992) tagger (followed by an affix-based tagger to try to annotate tokens remaining unknown after running the Brill model). Both taggers were trained using 90% of Ancora (Taulé, Martí, and Recasens, 2008) and its PoS tags as the training set and the remaining 10% as the development set. As stated

1 http://grupolys.org/software/MIOPIA/

in Chapter 2, an additional challenge for Spanish word-category disambiguation is that the use of accents is commonly ignored by people when writing in a web environment. To improve practical performance of our tagger, we have expanded the training set as explained in §2.2. We evaluated both the regular tagger and the tagger trained with the expanded set (sentences were copied without including any acute accent). We obtained an accuracy of 95.86% and 95.71% on the test set (which was not expanded), respectively, but we have observed that the regular tagger performs poorly on web texts. We hypothesize this is due to the fact the Ancora corpus is correctly written, which is not the case of the majority of the web reviews. However, we have observed that our cloned tagger was able to tag these type of reviews correctly (Example 8 in Chapter 2 was a real output of a regular and an expanded-set tagger on a sentence of the sentiment corpus we are using in this chapter).

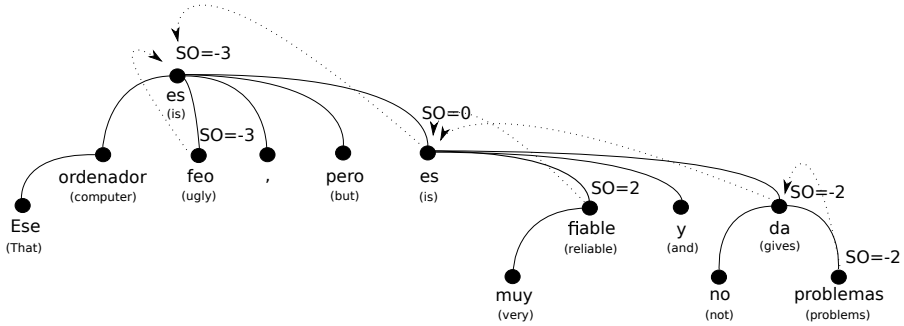
Once these steps have been performed, we use dependency parsing for analyzing the syntactic structure of each given sentence. In particular, we have used MaltParser and the Ancora corpus (same splits that the ones used to train the PoS tagger) to train a dependency parser based on the *Nivre arc-eager* algorithm (Nivre, 2008b). We achieved a LAS of 81.79% and a UAS of 86.76%, which is a competitive accuracy for Spanish. The best-performing system among the 19 participants in the CoNLL-X shared task (Buchholz and Marsi, 2006) reported a LAS of 82.25% and a UAS of 86.05% (note that, since that task used different training and test corpora, this should be taken as a rough indicator of performance and not as a direct comparison between parsers). This means that we have a solid base from which to reliably detect relevant syntactic phenomena like intensification, subordinate adversative clauses and negation; and misdetections are likely to be infrequent enough to not have a large impact in our system's performance. A more precise estimation of this impact could be obtained by task-oriented evaluation, but this would require a costly manual annotation process (Volokh and Neumann, 2012).

4.1.1 *Baseline*

To measure the impact of defining syntax-based rules, we first define an equivalent implementation to a purely lexical approach, where we calculate the SO of a sentence just taking into account common subjective nouns, adjectives, adverbs and verbs stored in a subjectivity lexica (in the case of this chapter, the dictionaries from Brooke, Tofiloski, and Taboada (2009)). The SO of each word spreads recursively to the upper levels of the dependency tree until ROOT is reached. Each head node aggregates the sentiment of its children. Syntactic constructions such as negation, subordinate adversative clauses or intensification are not considered at this time, to show the drawbacks of these kind

of simplistic sentiment analysis models. We exemplify this below these lines:

EXAMPLE 13 (Analysis of SSSA over a sentence by simply summing individual SO's). We draw a sentiment analysis on the dependency tree of the sentence *'Este ordenador es feo, pero es muy fiable y no da problemas'* (*'That computer is ugly, but it is very reliable and doesn't give problems'*) by only summing the semantic orientation of individual subjective words.



The sentence in the example is generally perceived as slightly positive, but this initial proposal classifies it as negative, because there are syntactic constructions that have been not considered in the base system, such as the negation *'no'* (*'not'*), the intensification *'muy'* (*'very'*) or the adversative subordinate conjunction *'pero'* (*'but'*).[2] In the following examples we describe how we deal with them and how we include these valence shifters on our approach.

□

4.1.2 *Intensification*

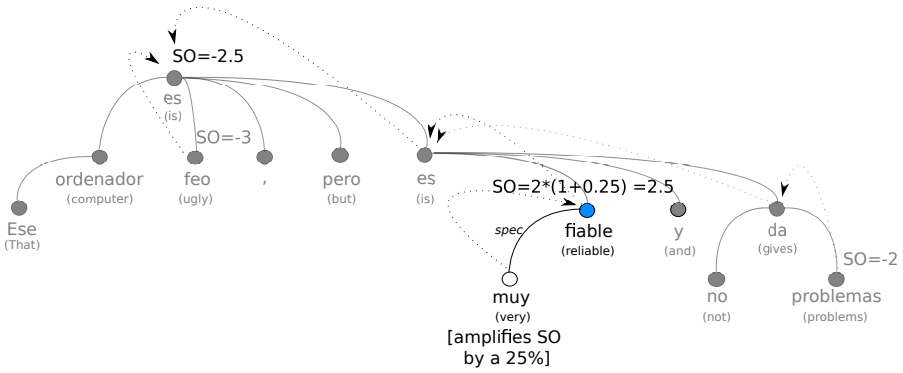
An intensifier is a word or an expression which plays the role of a valence shifter in a sentence. There are two types according to their category: *amplifiers* and *downtoners*. The former maximize semantic orientation of one or more tokens, such as *'muy'* (*'very'*); whereas the latter decrement it, e. g. *'en absoluto'* (*'not at all'*) or *'poco'* (*'little'*).

In some respects, our treatment of intensification is similar to that of Taboada et al. (2011), in the sense that amplifiers and downtoners are modeled as SO modifiers. Each intensifier has an associated percentage, positive if it is an amplifier and negative if it is a downtoner. However, ambiguous cases might appear where such lexical heuristics are not sufficient. For example, *'huge'* can be a subjective adjective introducing its own SO (e.g. *'The house is huge'*), but also an amplifier when it modifies a subjective noun or adjective (e.g. *'I have huge problems'*, where it makes *'problems'* more negative).

2 Throughout the chapter we will use italics to represent all the linguistic aspects that can shift the sentiment of a sentence

Syntax-based rules help overcome this problem without the need of window-based heuristics. In the case of Spanish and Ancora trees, whenever an adverb is a dependent of a specifier (dependency labels *spec* and *espec*) or an adjunct (dependency labels *cc* and *sadv*) type, we take that word as a valence shifter and its head as the exact scope to be shifted. Example 14 illustrates how SSSA manages intensification on the running example.

EXAMPLE 14 (Analysis of SSSA over a sentence when incorporating a treatment of intensification). We illustrate the effect on the sentiment calculation over the dependency tree of the sentence *'Este ordenador es feo, pero es muy fiable y no da problemas'* once the treatment of intensification is incorporated. We take *'fiable'* (*'reliable'*) as an intensified word, because its dependent node is an adverb and it is labeled with the dependency type *spec*. To calculate the sentiment of this piece of the sentence, we retrieve the original SO of *'fiable'*, which is 2, and we increase it by 25%, the percentage associated to the amplifier *'muy'* (*'very'*): $2 * (1 + 0.25) = 2.5$. Also, it is possible to nest the effect of two or more intensifiers to shift the SO of a term. Nested intensifiers are labeled with the *spec* dependency type and their head node is always another intensifier. In this case, we calculate the final valence shift by aggregating the percentages associated to different intensifiers, subsequently applying the resulting percentage to a token. For example, in *'en absoluto muy fiable'* (*'not very reliable at all'*), where *'en absoluto'* (*' not at all'*) has an associated percentage of -100%, we would calculate the semantic orientation of that expression as $2 * (1 + (0.25) + (-1)) = 0.5$.



□

Finally, there are other ways of emphasizing an idea. Exclamation marks make it possible to indicate a stronger conviction or a salient word in a sentence. For treating this phenomenon, we included *'!'* in the dictionary of intensifiers with a percentage value of +50% and we added the *f* dependency type (used for punctuation marks) to the algorithm for detecting intensified expressions.[3]

3 Unlike English, Spanish uses *'¡'* to begin exclamatory sentences, but it is customary to omit it in a web environment, and for this reason it has not been considered here.

4.1.3 *Subordinate adversative clauses*

A subordinate adversative clause expresses an event or fact that is the opposite to that of the main clause. In an SA context, we hypothesize that these type of constructions are a way of restricting, excluding or amplifying the sentiment reflected by both the main and subordinate clauses. We consider subordinate adversative clauses as a special case of intensification, but involving clauses, not individual terms. For example, the sentence *'The actor acted badly but the movie was great'* is perceived as slightly positive because the conjunction *'but'* implicitly gives more importance to the subordinate adversative clause *'the movie was great'*, while the main clause is partially ignored.

Type of conjunction	**Weight for main clause**	**Weight for subordinate clause**
Restrictive	0.75	1.4
Exclusive	0	1

Table 7: Weights of restrictive and exclusive conjunctions

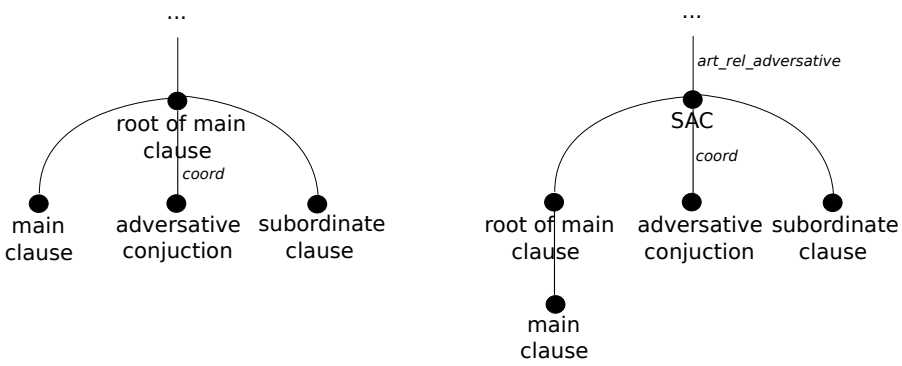
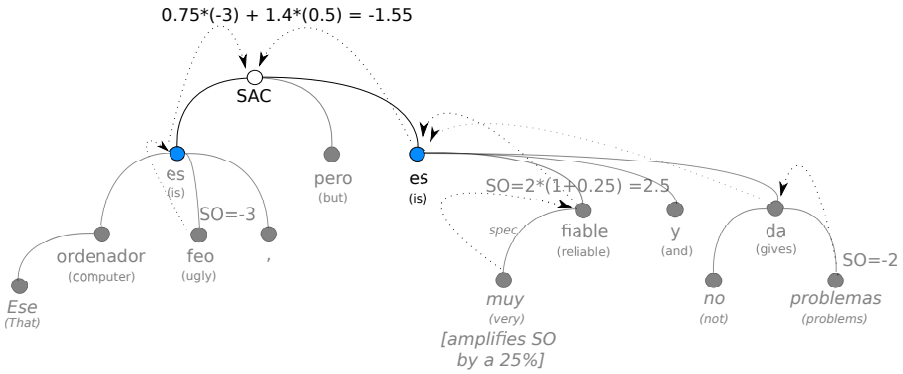


Figure 2: Display of the reorganization of subordinate adversative clauses on Ancora trees to be processed by SSSA

In this respect, we distinguish two different types of adversative conjunctions, as is pointed out in Campos (1993), Chapter 3. The first type, *restrictives*, increase the sentiment of the subordinate clause and decrease the so of the main clause. The second type, *exclusives*, ignore totally the sentiment reflected in the main clause. Unfortunately, the Ancora corpus uses different dependency trees and dependency types for representing different adversative clauses. In this work, we only treat sentences that are uniformly structured: we take *'pero'* (*'but'*) and *'mientras'* (*'while'*) as restrictive conjunctions and *'sino'* (*'but rather'*) and *'sino que'* (*'but on the other hand'*) as exclusives.

Table 7 illustrates how we weight both types.[4] In order to homogenize in the future all syntactic representations of the subordinate adversative clauses, we carried out a reorganization of dependency trees, as shown in Figure 2. Moreover, it simplifies our SO calculation algorithm to weight both the main and subordinate clauses. For this purpose, we include an artificial node, called SAC, at the top of subordinate adversative clauses; and a new dependency type, *art_rel_adversative*, to identify syntactically the beginning of this type of clause. In Example 15 we show the effect of treating this linguistic phenomenon on the running example.

EXAMPLE 15 (Analysis of SSSA over a sentence when incorporating a treatment of subordinate adversative clauses). The dependency tree below shows the reorganization of the sentence *'Este ordenador es feo, pero es muy fiable y no da problemas'* and how we calculate the sentiment of a sentence once the treatment of adversative subordinate clauses is incorporated.



Thus, our sentiment analyzer would identify an artificial node, would decrease the SO accumulated in the main clause by 25% (multiplying by 0.75) and amplify by 40% (multiplying by 1.40) the sentiment of the subordinate sentence: $0.75 * (-3) + 1.40 * (0.50) = -1.55$.

□

4.1.4 *Negation*

Negation is one of the most challenging phenomena to handle in SA, since its semantic scope can be non-local (e.g. *'I do not plan to make you suffer'*). Existing unsupervised lexical approaches are limited to considering a snippet to guess the scope of negation. Thus, it is likely that they consider as a part of the scope terms that should not be negated from a semantic point of view.

The most common and simple way to negate a sequence of tokens in Spanish is the adverb *'no'* (*'no'*/*'not'*), but other terms such as

4 The weights have been empirically established over the SFU Spanish review corpus. We tested values between 0 and 2 both for main and subordinate clauses using steps of 0.15 and 0.2, respectively.

'sin' (*'without'*) or *'nunca'* (*'never'*) are frequently employed. However, some types of Spanish sentences usually require the use of double negatives to make a negative sentence.

In this respect, words like *'nada'* (*'nothing'*), *'ninguno'* (*'none'*) or *'nadie'* (*'nobody'*) are commonly preceded by *'no'*. Moreover, the difference between a negating term and a downtoner is diffuse. Tokens like *'apenas'* (*'barely'*) or *'casi'* (*'almost'*) could easily be classified in either of these two categories. We have chosen to consider these type of expressions as intensifiers and therefore we only consider explicitly as negators the adverbs *'no'*, *'nunca'* and *'sin'*, which cover a great number of negative sentences. Our treatment of a negation consists of two basic steps: 1) *identify the scope of a negating term* and 2) *modify the semantic orientation of affected tokens*.

Scope identification

The procedure for identifying the scope of a negation depends on the adverb used in the phrase.

The syntactic structure used in Ancora for representing an adverb *'sin'* assures us that its child node should be the scope of negation, without needing to analyze the dependency type. But we cannot assume the same for the negators *'no'* and *'nunca'*. Usually they are represented as leaf nodes and the candidate scope of negation always involves a head node or a collection of sibling nodes, so we require a more complex algorithm for their treatment. We use a procedure based on Jia, Yu, and Meng (2009), which uses a parse tree and a collection of special rules to identify the scope of each negation. Firstly, the candidate scope for a negator is identified. Then, the exact scope is determined by searching *delimiters* by means of a syntactic heuristic procedure. A delimiter is a token that has the capability to eliminate some words from the candidate scope of a negating term. We have adapted this procedure to profit from the additional information provided by the syntactic structure of the sentence. We use dependency types to directly extract the exact scope without identifying delimiter words. When a token has a negator *'no'* (*'not'*) or *'nunca'* (*'never'*) as a child node and it is a dependency of type *'neg'* or *'mod'*; we try the collection of syntactic heuristic rules shown in Figure 3, in the following order:[5]

1. *Subjective parent rule*: Whenever a parent node of a negating term has sentiment, only that node is negated. Figure 3.a shows how we take the scope when this rule matches. For example, in the sentence *'he does not praise my work'*, the negation *'not'* depends on *'praise'*, which is included as a subjective word in the so dictionaries, so we consider this term as the scope of the negation.

5 Only the first matching rule is applied.

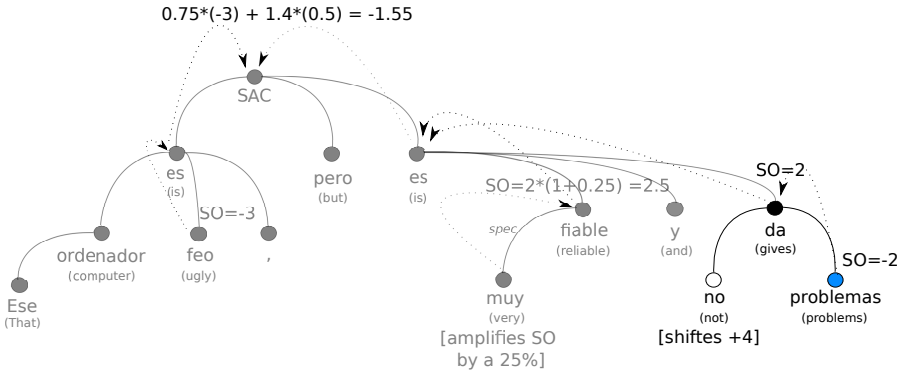
2. *Subject complement/Direct object rule*: Whenever a branch at the same level as a negation node is labeled with a dependency of type subject complement (*atr*) (e. g. *'the meal is not good'*) or a direct object (*cd*) (e. g. *'the meal does not look good'*), our sentiment analyzer negates that branch, as we show in Figure 3.b.

3. *Adjunct rule*: Whenever a negating term has an adjunct branch (*cc*) at the same level, the sentiment of that branch is shifted. If there is more than one adjunct, only the first one is negated, as shown in Figure 3.c. For example, in the sentence *'he does not work efficiently on Fridays'*, our method takes the mood adjunct (*'efficiently'*) as the scope of the negation, because it is the nearest to the negation.

4. *Default rule*: Figure 3.d shows how when none of the previous rules matches, we consider as scope the sibling branches of a negator.

We now explain in more detail the treatment of negation in the running example.

Example 16 (Analysis of sssa over a sentence when incorporating a treatment of negation). The figure draws over the dependency tree how the so of the scope of negation, which is *'problemas'* (*'problems'*), is modified by this amount. The word *'problemas'* has a so of -2, and the phrase *'no da problemas'* has a so of $-2 + 4 = 2$.



□

In Example 16 we can see that the word *'no'* has as its head the verb *'da'* (*'gives'*). Our method first tries to apply the *subjective parent rule*, but in this case, this is not a subjective node, so that rule is ignored. Then, our procedure continues with the *direct object rule*, which matches, because there is a direct object dependent (identified by *cd*) at the same level as the negation, so this rule is applied and takes *'problemas'* (*'problems'*) as the scope of negation.

Polarity flip

There are several ways of taking into account the effect of negation.

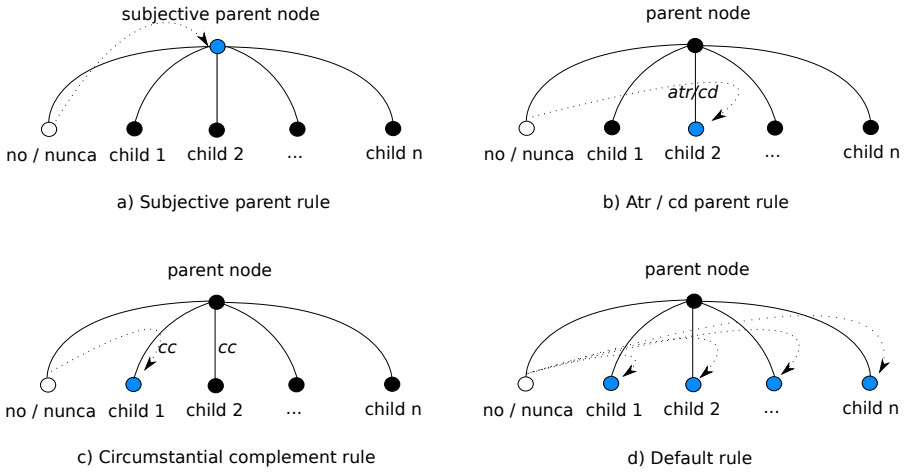


Figure 3: Display of the heuristic rules used by SSSA to identify the scope of negating terms

On the one hand, machine learning methods can explicitly handle this by unifying the negator and the negated word into a single feature (Sidorov et al., 2013), using two-steps model to identify negation cues and their scope (Cruz, Taboada, and Mitkov, 2015) or modeling recursive neural networks over dependency trees (Socher et al., 2012).

On the other hand, the simplest way to negate a word in semantic approaches is to invert the SO (e. g. if SO (good) = 2 then SO (not good) = -2). The main drawback of this method is that it is not coherent with human intuition. For example, if the SO of *'fascinating'* is 5 the sentiment of *'not fascinating'* would be -5, when it could even be considered a slightly positive expression.

Our polarity flip algorithm follows a shift negation method where the SO value is shifted toward the opposite polarity by a fixed amount: following Taboada et al. (2011), we have chosen a flip value of 4 for the adverbs *'no'* (*'not'*) and *'nunca'* (*'never'*). Example 16 shows how this process works. For the adverb *'sin'* (*'without'*), based on our experimental setup, we have chosen a value of 3.5. We hypothesize this kind of negation as being less potent, given that its scope is fairly local. Experimental results described below show an improvement in accuracy when carrying out this strategy.

4.1.5 *Adding lexical functionalities*

Along with the syntactic issues, there are other factors that can influence the overall sentiment, such as the discourse structure (Pang and Lee, 2008). The order in which authors express their opinions can change the sentiment polarity. It is customary that the final sentences of a text play the role of a summary or conclusion, giving implicitly more emphasis to this part of the document. To simulate this phenomenon, our proposal increases the sentiment of the last three

sentences of a given review. We chose a value of 0.75 based on experimental evidence. Also we note that by increasing the SO of the nouns, adjectives, verbs and adverbs from Brooke, Tofiloski, and Taboada (2009) by 20% our approach improved the performance on our development corpus. Thus, the SO considers values between -6 and 6. This modification is applied both to the hand-created dictionaries and to the automatically enriched dictionaries explained in Section 4.2. All the strategies that improved the performance of our proposal were included in our final version.

The motivation of all these optional features was experimental, taking the SFU Spanish review corpus as the development set, but they also work satisfactorily on other long text corpora, as we show in Section 4.3.2.

4.2 DOMAIN ADAPTATION

The generic SO of dictionaries can be inadequate in a particular domain. Entertainment contexts are some of the typical fields where this phenomenon occurs more frequently (e. g. words such as *'killer'* or *'horror'* should not be clear negative indicators if we are discussing about movies). In this section, we provide a semi-automatic method to adapt and enrich semantic dictionaries to a specific area and we use CorpusCine, a corpus of Spanish movie reviews, as an example. In Section 4.3.1 we detail the content of this corpus. In Section 4.3.2 we illustrate how our adaptation method improves the performance for this domain.

Our aim is to learn the polarity of subjective words in a given domain. This implies discovering words which are not present in the generic dictionary and also adapting the polarity of words already present in the dictionary to their use in the specific field.

For the first task, we learn which tokens are good polarity classifiers in the area in question by extracting the most representative words in that domain relying on their *information gain* (IG) (Hall et al., 2009; Mitchell, 1997) with respect to the classes (in our case we only have positive and negative classes). Once we have classified the attributes, we need to give an SO to each selected word. We hypothesize that if an attribute appears more frequently in positive than in negative texts, that feature must be positive, and vice versa. If a word is positive we calculate its SO with the equation (7), and if it is negative we employ equation (8).

$$\mathrm{OS}_{word_i} = \frac{\log_2(\frac{x_i+1}{y_i+1})}{\log_2 z} \times 5 \tag{7}$$

$$\mathrm{OS}_{word_i} = \frac{\log_2(\frac{y_i+1}{x_i+1})}{\log_2 w} \times -(5+\alpha) \tag{8}$$

where:

- x_i represents the number of positive texts where $word_i$ appears, and y_i the number of negative texts.

- z represents the maximum value x_i/y_i for all i.
- w is the maximum coefficient y_i/x_i, for all i.
- α is a weight factor given to negative words.

The resulting values are normalized between 5 and -5, in order to make them comparable with the values in Brooke, Tofiloski, and Taboada (2009) dictionaries. The words with an so close to 0 will represent neutral terms. We need to create a *pessimistic* dictionary to improve performance and counteract the *optimistic* tendency of CorpusCine, a characteristic widely explained in other studies, as we will show in Section 4.3.2, and the reason why parameter α is used in Equation 8. A possible option to do this is to increase the semantic orientation of negative words. Another equivalent option consists of including more negative than positive words. Both perspectives will be analyzed in Section 4.3.2.

After creating the domain dictionary, we must merge it with the generic dictionary. We hypothesize that if the so is less than 0.5 in absolute value, the word is not a clear subjective word, so we discard it. For the rest of the words in the domain dictionary, we check the generic semantic orientation of the word in Brooke, Tofiloski, and Taboada (2009) dictionaries. If it does not have a generic so specified in that dictionary or it has a different sign than the domain specific so obtained, the latter prevails. If both the generic and the adapted so have the same sign, then the generic so prevails. This means that our method will only change the so of words that are clearly used with a non-standard polarity in the target domain, but it will not try to adjust the exact so value for words where the obtained sign matches the one in the dictionary. As an example, Table 8 shows the top five representative informative attributes in the movie domain while Table 9 shows some words of the movie domain that have changed their semantic orientation with respect to the general dictionary.

4.3 EXPERIMENTS

4.3.1 *Datasets*

We used three annotated corpora:

- The SFU Spanish Review Corpus (Brooke, Tofiloski, and Taboada, 2009) is a collection of 400 Spanish reviews on cars, hotels, washing

Ranking	Word	Generic so	Movie domain so
1	*'perfecto'* (*'perfect'*)	4	1.808
2	*'obra'* (*'work'*)	5	1.139
3	*'maestro'* (*'masterly'*)	0	1.760
4	*'imprescindible'* (*'indispensable'*)	4	3.259
5	*'peor'* (*'worse'*)	-2	-1.712

Table 8: Top 5 discriminative tokens in the CorpusCine (film domain) according to information gain

Word	Generic so	Movie domain so
'violencia' (*'violence'*)	-5	1.511
'guerra' (*'war'*)	-2	1.310
'zombi' (*'zombie'*)	-1	0.730
'kryptonita' (*'kryptonite'*)	0	-1.981
'bestseller'	4	-1.250

Table 9: Generic vs. adapted so's to the film domain

machines, books, cell phones, music, computers, and movies from the `www.ciao.es` web site. Each category has a total of 25 favorable and 25 unfavorable reviews. As usually happens in reviewing web sites, texts have unstressed words, unrecognized abbreviations and ungrammatical sentences. This allows us to evaluate our proposal in a real and complex environment. Moreover, The Spanish SO-CAL was developed on this corpus, so their lexicon-based approach and our dependency parsing-based method can be compared.

• CorpusCine reviews (Cruz Mata, 2011) is a collection of 3 878 movie reviews written in Spanish from the `www.muchocine.net` web page. Each document is rated between one and five stars, where one is the most negative rating and five the most positive. There are 351 one-star, 923 two-star, 1 253 three-star, 890 four star and 461 five-star reviews. We classify one or two-star documents as negative. Three-star reviews are discarded because we consider them as neutral or mixed reviews. This is a widely accepted strategy that has been employed in other studies Cruz Mata (2011) and corpora, like the SFU Spanish review corpus[6]. Documents ranked with four or five stars are taken as positive reviews.

• HOpinion[7] is a collection of 17 934 hotel reviews extracted from `www.tripadvisor.es`, rated between one and five stars. There are 841 one-star, 1 269 two-star, 3 468 three-star, 6 244 four-star and 6 112 five-

6 This issue is detailed on the readme file of `www.sfu.ca/~mtaboada/download/downloadCorpusSpa.html`

7 `http://clic.ub.edu/corpus/hopinion`

star reviews. We followed the same strategy as in CorpusCine to evaluate it, discarding three-star texts.

4.3.2 *Evaluation*

Results on SFU Spanish Review Corpus

Table 10 shows the performance of our system with a number of different options on the SFU Spanish review corpus. All features contribute to performance. One of the most important improvements in accuracy comes from the treatment of negation. As we can see, before incorporating this feature our approach favors positive classifications. This likely happens as the result of a human tendency to positive language (Kennedy and Inkpen, 2006). People usually negate positive sentences to express an unfavorable opinion. For example, it is common to use expressions like *'not good'* instead of *'bad'* or *'I don't like it'* instead of *'I dislike it'*. Even after processing negating terms, a lexicon-based system such as the English SO-CAL increases the final SO of any negative expression by 50% to overcome that positive bias, improving its performance by around 6% with this strategy. However, in our current implementation that feature gave no benefit. This suggests to us that our negation algorithm performs well, at least in a general context.

Category	$\mathbf{r}_{neg}$	$\mathbf{r}_{pos}$	**Accuracy**
Baseline	0.310	**0.925**	0.618
+intensification	0.450	0.870	0.660
+adversative clauses	0.455	0.885	0.670
+negation	**0.745**	0.765	0.755
Final proposal	0.740	0.830	**0.785**

Table 10: Performance of SSSA on the SFU Spanish Review Corpus with a variety of options enabled

Table 11 shows the performance of our final approach on each subcorpus of the SFU Spanish review corpus. As we can see, there are significant differences in performance depending on the category. For domains where quality criteria are reasonably objective, such as hotels, computers or washing machines, our proposal performs well (over 80% accuracy), because the generic SO is usually adequate. But the same is not true for entertainment domains such as movies, books and music, where performance falls below the average. We believe this is mainly due to the problem of generic semantic orientations, as we have discussed throughout the chapter, which primarily affects this type of domains. Moreover, movies or books are contexts where personal tastes are particularly important. For example, the fragment

'is a low-budget movie' is in principle a negative sentence, but it could be positive for a person who loves B movies. This makes it difficult to assign a semantic orientation according to the sentiment of users, even for a particular domain.

Category	$\mathbf{p}_{neg}$	$\mathbf{r}_{neg}$	$\mathbf{p}_{pos}$	$\mathbf{r}_{pos}$	**Accuracy**
Hotels	0.88	**0.92**	**0.88**	0.88	**0.90**
Computers	**0.91**	0.80	0.82	**0.92**	0.86
Washing machines	0.79	0.88	0.86	0.76	0.82
Cell phones	0.86	0.72	0.76	0.88	0.80
Cars	0.77	0.68	0.71	0.80	0.74
Music	0.84	0.64	0.71	0.88	0.76
Books	0.80	0.64	0.70	0.84	0.74
Movies	0.67	0.64	0.76	0.68	0.66

Table 11: Performance of SSSA per category on the SFU Spanish Review Corpus

Table 12 compares the performance of various methods on the SFU Spanish review corpus. Our syntactic proposal improves the accuracy of The Spanish SO-CAL by about 6%, even though the SO-CAL is a system with more functionality (e. g. treatment of irrealis). This suggests that parsing is useful in order to resolve the polarity of a given text. In particular, we believe that an effective treatment of negation requires a more complex algorithm than a purely lexicon-based technique.

We also compare our proposal with an ML method. More specifically, we have trained a SVM as a classifier. We have relied on WEKA to build it, using libSVM (Chang and Lin, 2011). Specifically, we chose an SVM of type C-SVC, a radial basis function as the kernel type and a value of 1 for the cost parameter. Testing was done with 10-fold cross-validation. Data was preprocessed in order to change the words to their lowercase form, and we have employed the output word counts as the weighting factor. Over the SFU Spanish review corpus, our syntax-driven analyzer provides better accuracy than the SVM , reinforcing the idea that the ML approach is not the best technique to build a general domain polarity classifier, at least when performing a binary classification[8]. Finally, we tested a hybrid approach, labeled on Table 12 as 'SVM + our SO as feature': we analyzed each text with our proposal and we included the SO obtained as a feature for the SVM

8 We tested various configurations with different weighting factors and different types of preprocessing, but we only show the configuration who achieved the best performance. Results are similar to the ones presented on the same corpus by Brooke, Tofiloski, and Taboada (2009).

Method	r_{neg}	r_{pos}	Accuracy
SSSA	0.7400	**0.8300**	**0.7850**
SVM + our SO as feature	**0.7490**	0.7700	0.7594
The Spanish SO-CAL			0.7425
SVM	0.7230	0.7270	0.7250

Table 12: Performance on the SFU Spanish Reviews corpus (SSSA vs. other methods)

Method	r_{neg}	r_{pos}	Accuracy
SVM	0.5800	**0.9930**	**0.9328**
SSSA	**0.7294**	0.9218	0.8938
SVM $_{sfu}$	0.6770	0.7940	0.7766

Table 13: Performance on the HOpinion corpus (SSSA vs. various methods)

. However, the resulting accuracy was worse than the one obtained with our system alone.

Results on HOpinion

Table 13 shows the performance on HOpinion. Results are similar to those obtained on the hotel category of the SFU Spanish review corpus, achieving an accuracy of 0.8938.

We also built an SVM classifier specific to HOpinion, applying lemmatization to the texts, using TF-IDF[9] as weighting factor and selecting the default configuration of WEKA for the SVM (type C-SVC, a radial basis function as the kernel type and 1 as the cost parameter). We used 10-fold cross-validation to evaluate it, achieving an accuracy of 0.9328. This supervised classifier did not satisfactorily learn negative reviews due to the low number of unfavorable opinions in the corpus. Finally, we evaluated an SVM trained with the SFU Spanish Review corpus (SVM_{sfu}) on HOpinion. In this case, we did not apply lemmatization, as we did in the classifier trained on the SFU Spanish Reviews, and we changed each word to its lowercase form and used their total output count as the weighting factor.

Results on CorpusCine

Table 14 shows the performance on CorpusCine obtained by the different approaches explained. Moreover, we included the results obtained by a supervised approach presented in Cruz Mata (2011). This specific domain method uses five morphosyntactic patterns to extract sentiment bigrams using multiple seed words (Turney, 2002) to then

9 We tested other weighting factors such as the binary or the total occurrence of each term, but we achieved the best performance using TF-IDF.

Method	r_{neg}	r_{pos}	Accuracy
SVM	**0.8440**	**0.8220**	**0.8328**
SSSA with domain adaptation	0.7997	0.8024	0.8011
Cruz Mata (2011)[10]	0.8250	0.7250	0.7750
Our proposal	0.4804	0.7935	0.6415
SVM $_{sfu}$	0.6250	0.6130	0.6179

Table 14: Performance on the CorpusCine corpus (SSSA vs. various methods)

Polarity	Number of stars	sssa	sssa with domain adaptation
Negative	1	0.6923	**0.9003**
	2	0.3948	**0.7614**
Positive	4	**0.7933**	0.7674
	5	0.7939	**0.8698**

Table 15: Accuracy per star score on the CorpusCine corpus for SSSA with generic and adapted semantic orientation lexica

calculate their SO. It provides a supervised technique which uses an optimal threshold for categorizing favorable and unfavorable texts.

We also built an SVM classifier specific to CorpusCine.

Also, We used tf-idf as the weighting factor. We selected the default configuration of WEKA for the SVM (type C-SVC), a radial basis function as the kernel type and 1 as the cost parameter). We used 10-fold cross-validation to evaluate it. Moreover, as we did with HOpinion, we evaluated an SVM trained with the SFU Spanish Review corpus (SVM_{sfu}) on CorpusCine. The performance drops below our generic approach, which reinforces the idea that ML methods are highly domain dependent. In contrast, our generic proposal shows a performance similar to that obtained on the *'movie'* category of the SFU Spanish review corpus, a result that confirms the domain independence of the proposal.

Finally, to test our proposal with dictionaries adapted to the movies domain we have also used 10-fold cross-validation. For each fold we extracted around 22 000 attributes (there are many more positive than negative attributes) from WEKA and for each one we built a dictionary using the training set, and we tested it against the development set.

As we can see, our adapted approach improves over the performance obtained with our generic approach by about sixteen percentage points. Moreover, we neutralize the positive bias that our generic system presented on CorpusCine. Table 15 compares, in greater detail, the performance of our proposal on CorpusCine, before and after adapting it to the movie domain.

We have observed that unfavorable reviews had a high presence of condescending and ironic expressions, complicating the semantic analysis of those texts. To overcome this, we chose to build a dictionary where negative words had more relevance. Figure 4 shows how different weightings for the negative words (the parameter α explained in equation 8), and the different number of positive and negative entries in our specific semantic movie dictionary, affect the performance. We identify each graphic with a notation P-N, where P means that for that case of study we have only considered the first P percent of the positive attributes extracted from WEKA and the first N percent of the negative ones. For example, *75-25* would represent a case where we only employed 75% of the best positive classifiers and only the first 25% of the negative ones. Note that for each weight, the number of negative words is different, because with a higher negative weighting there are more negative words with an SO greater than 0.5 in absolute value, the threshold value established in §4.2. Below we provide a brief explanation for each graphic included in Figure 4:

- *10-10*: The improvement in performance is minimum. Most of the words included in the dictionary are already present in Brooke, Tofiloski, and Taboada (2009) and have the same polarity, so our system takes few words from the specific domain dictionary.

- *75-25*: This configuration does not work well, due to the optimistic trend to favor positive classifications that our initial proposal presents in this particular corpus.

- *50-50*: The behavior is similar to that explained in the previous point (75-25). Although we employ 50% both for positive and negative words, the dictionary extracted from WEKA has many more positive attributes, so this is also a configuration that favors positive classification. However, we can see how by increasing negative weightings we obtain a good final performance.

- *25-75*: With this setup we obtain a good baseline, but performance decreases when we employ high negative weightings, because our system becomes too favorable to negative classifications.

- *95-100*: This was the best setup. We achieved an accuracy of 0.8011 with a negative weighting factor of 5.5.

- *100-100*: As in the 50-50 configuration, with large negative weightings we can obtain a high performance and counteract the optimistic tendency.

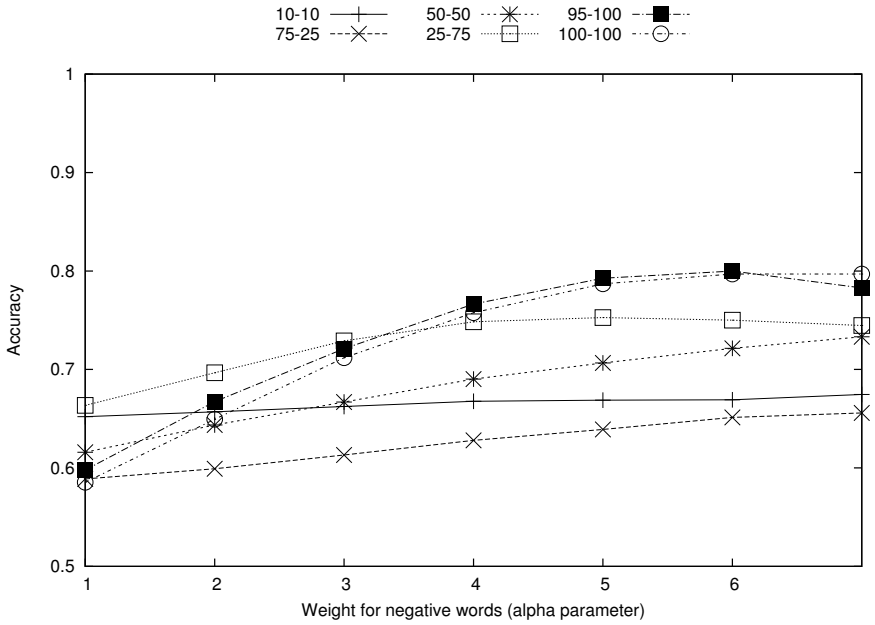


Figure 4: Accuracy on CorpusCine increasing negative word weighting

4.4 CONCLUSION

In this chapter, we have described a syntax-based method for sentiment analysis of Spanish reviews. We used dependency-based methods to treat some relevant linguistic aspects in SA, such as intensification, subordinate adversative clauses and negation. Two sets of experiments were performed to compare our method to other existing techniques. Experimental results on a general domain corpus show that our syntactic proposal improves ML and lexicon-based approaches on a general-domain setting. Moreover, we performed an evaluation over a specific domain corpus (movies), where ML techniques obtain a much better baseline accuracy than semantic approaches, due to the invalidity of the generic semantic orientations. We have proposed a semi-automatic method to enrich and adapt the semantic dictionaries to a particular field, and we have applied it to our model. Experiments show a good performance, obtaining an accuracy close to that of ML classifiers and improving over other existing domain specific systems.

5

A PROPOSAL TO UNIVERSAL, UNSUPERVISED SYNTAX-BASED SENTIMENT ANALYSIS

In previous chapters, we have developed two methods for determining the polarity of both short and long texts written in Spanish. In Chapter 3 we adapted a purely lexicon-based and multilingual system, SentiStrength, to Spanish. Its main weakness is the incapacity of defining rules to handle phenomena involving non-local semantic compositions, since it just relies on window-based rules. To address this challenge, in Chapter 4 we proposed a syntactic method, which bases its analysis on the trees returned by a dependency parser, but unfortunately it was heavily dependent, not just on the language, but also on the criteria used to annotate dependency structure. In this chapter, we formalize the approach presented in Chapter 4, by introducing a formalism for compositional operations, allowing the creation of arbitrarily complex rules to tackle relevant phenomena for SA, for any language and syntactic dependency annotation, so we can handle multilinguality as easily as SentiStrength (Chapter 3) does. The main contribution is the introduction of the first universal syntax-based model for compositional sentiment analysis.

For this purpose, we implement and evaluate a set of practical universal operations defined using part-of-speech tags and dependency types under the universal guidelines of Petrov, Das, and McDonald (2011), McDonald et al. (2013) and Nivre et al. (2016): universal annotation criteria that can be used to represent the morphology and syntax of any language in a uniform way.

We first build different monolingual models that share the same compositional operations across different languages (English, Spanish and German). The approach outperforms existing unsupervised approaches as well as state-of-the-art compositional supervised models (Socher et al., 2013) on domain-transfer settings, and shows that the operations can be shared across languages, as they are defined using universal guidelines. We will be referring to our system as universal, unsupervised, uncovered sentiment analysis (UUUSA).

We then build a single multilingual model that in addition shares the subjectivity lexica and the tagging and parsing models, apply it to the context of five official languages of the Iberian Peninsula (Spanish, Portuguese, Basque, Catalan and Galician) and show its robustness when we are using a single multilingual pipeline in an end-to-end application.

UUUSA can be downloaded from: `http://grupolys.org/software/UUUSA/`

5.1 DESCRIPTION

Detecting the scope of non-local linguistic phenomena in NLP is one of the applications where parsing can be useful, as we showed in Chapter 4. However, the system there presented (SSSA), was not just monolingual (only intended for Spanish), but also dependent on the annotation of the training treebank, and so the rules were annotation-dependent too. Consequently, the sets of rules proposed was also annotation-dependent and complicate adaptation, especially in multilingual environments (or even within the same language, since different treebanks might follow different guidelines).

In this chapter we address this challenge by proposing a formalism for *compositional operations* (§5.1.1), allowing the creation of arbitrarily complex rules to tackle relevant phenomena for SA, for any language and syntactic dependency annotation. The formalism is independent on the treebank, but for the sake of practice and to reinforce the utility of the system in multilingual environments, we are using universal treebanks that will allow us to define operations that can be used across different languages, without changing them in any way. These are briefly discussed now in §5.1.

Universal Dependency Treebanks

A number of dependency treebanks for a variety of languages have been made available in the last years (Buchholz and Marsi, 2006). However, the problem on creating truly multilingual syntax-based NLP systems persists, since such treebanks are heterogeneous, relying on different schemes and guidelines. Such differences might be superficial or deep, but all of them complicate creating a single multilingual framework to solve a particular task. In this respect, McDonald et al. (2013) tackled this challenge and proposed a new collection of treebanks with homogeneous syntactic dependency annotations, that have been revised through the years (Nivre et al., 2016).

By taking advantage of these treebanks, in this chapter we implement and evaluate a set of practical universal compositional operations for SA and that can be shared across different languages.

5.1.1 *Compositional operations*

Prior to defining the concept of compositional operations, we introduce some additional functions that we will be using throughout this chapter: given a dependency tree $T = (V, E)$, and a node $i \in V$, we define a set of functions to obtain the context of node i:

- $ancestor_T(i, \delta) = \{k \in V$: there is a path of length δ from k to i in $T\}$, i. e. the singleton set containing the δth ancestor of i (or the empty set if there is no such node),

- *children*$_T(i) = \{k \in V \mid i \rightarrow k\}$, i. e. the set of children of node i,
- *lm-branch*$_T(i, d) = \min\{k \in V \mid i \xrightarrow{d} k\}$, i. e. the set containing the leftmost among the children of i whose dependencies are labeled d (or the empty set if there is no such node).

Our compositional SA system will associate an SO value σ_i to each node i in the dependency tree of a sentence, representing the SO of the subtree rooted at i. The system will use a set of compositional operations to propagate changes to the semantic orientations of the nodes in the tree. Once all the relevant operations have been executed, the SO of the sentence will be stored as σ_0, i. e. the semantic orientation of the root node.

A compositional operation is triggered when a node in the tree matches a given condition (related to its associated PoS tag, dependency type and/or word form); it is then applied to a scope of one or more nodes calculated from the trigger node by ascending a number of levels in the tree and then applying a scope function. More formally, we define our operations as follows:

DEFINITION 3. Given a dependency tree $T(V, E)$, a **compositional operation** is a tuple $o = (\tau, C, \delta, \pi, S)$ such that:

- $\tau : \mathbb{R} \rightarrow \mathbb{R}$ is a **transformation function** to apply on the SO (σ) of nodes,
- $C : V \rightarrow \{true, false\}$ is a predicate that determines whether a node in the tree will **trigger** the operation,
- $\delta \in \mathbb{N}$ is a number of levels that we need to ascend in the tree to calculate the scope of o, i. e. the nodes of T whose SO is affected by the transformation function τ,
- π is a priority that will be used to break ties when several operations coincide on a given node, and
- S is a scope calculation function that will be used to determine the nodes affected by the operation.

□

In practice, our system defines $C(i)$ by means of sets of words ($\{w_0, w_1, ..., w_n\}$), part-of-speech tags ($\{p_0, p_1, ..., p_m\}$) and/or dependency types such that the operation will be triggered if w_i, p_i and/or the head dependency of i are in those sets. Compositional operations where $C(i)$ is defined using only universal tags and dependency types, and which therefore do not depend on any specific words of a given language, can be shared across languages, as shown in §5.3.

We propose two options for the transformation function τ:

- *shift*$_\alpha(\sigma) = \begin{cases} \sigma - \alpha & \text{if } \sigma > 0 \\ \sigma + \alpha & \text{if } \sigma < 0 \end{cases}$ where α is the shifting factor and $\alpha, \sigma \in \mathbb{R}$.

• $weighting_\beta(\sigma) = \sigma \times (1 + \beta)$ where β is the weighting factor and $\beta, \sigma \in \mathbb{R}$. [1]

The scope calculation function, S, allows us to calculate the nodes of T whose so is affected by the transformation τ. For this purpose, if the operation was triggered by a node i, we apply S to $ancestor_T(i, \delta)$, i. e. the δth ancestor of i (if it exists), which we call the **destination node** of the operation. The proposed scopes are as follows (see also Figure 5):

• *dest* (*destination node*): The transformation τ is applied directly to the so of $ancestor_T(i, \delta)$ (see Figure 5.a).

• *lm-branch*^{d} (*branch of* d): The affected nodes are $lm\text{-}branch_T\ (ancestor_T(i, \delta), d)$ (see Figure 5.b).

• rc^n (n *right children*): τ affects the so of the n smallest indexes of $\{j \in children_T(ancestor_T(i, \delta)) \mid j > i\}$, i. e. it modifies the global σ of the closest (leftmost) n right children of $ancestor_T(i, \delta)$ (see Figure 5.c).

• lc^n (n *left children*): The transformation affects the n largest elements of $\{j \in children_T(ancestor_T(i, \delta)) \mid j < i\}$, i. e. it modifies the global σ of the closest (rightmost) n left children of $ancestor_T(i, \delta)$ (see Figure 5.d). [2]

• *subjr* (*first subjective right branch*): The affected node is $\min\{j \in children_T\ (ancestor_T(i, \delta)) \mid j > i \wedge \sigma_j \neq 0\}$, i. e. it modifies the σ of the closest (leftmost) subjective right child of $ancestor_T(i, \delta)$ (see Figure 5.e).

• *subjl* (*first subjective left branch*): The affected node is $\max\{j \in children_T\ (ancestor_T(i, \delta)) \mid j < i \wedge \sigma_j \neq 0\}$, i. e. it modifies the σ of the closest (rightmost) subjective left child of $ancestor_T(i, \delta)$ (see Figure 5.f).

Compositional operations can be defined for any language or dependency annotation criterion. While it is possible to add rules for language-specific phenomena if needed (see § 5.1.2), in this chapter we focus on universal rules to obtain a truly multilingual system. Apart from universal treebanks and PoS tags, the only extra information used by our rules is a short list of negation words, intensifiers, adversative conjunctions and words introducing conditionals (like the English "if" or "would"). While this information is language-specific, it is standardly included in multilingual sentiment lexica which are available for many languages, so it does not prevent our system from working on a wide set of languages without any adaptation, apart from modifying the subjective lexicon.

1 From a theoretical point of view, β is not restricted to any value. In a practical implementation, β values (which will vary according to the intensifier) should serve to intensify, diminish or even cancel the σ of the affected scope in a useful way. In this chapter, β's for intensifiers are directly taken from existing lexical resources and are not tuned in any way, as explained in §5.3.

2 lc^n and rc^n might be useful in dependency structures where elements such as some coordination forms (e.g. it is *'very expensive and bad'*) are represented as children of the same node, for example.

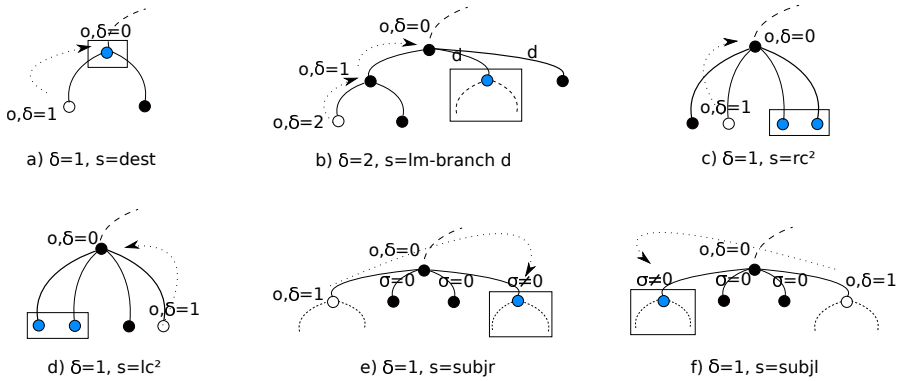


Figure 5: Graphical representation of the proposed set of influence scopes S. ◯ indicates the node that triggers an operation o, □ the nodes to which it is applied (colored in blue).

5.1.2 *An algorithm for compositional computation*

To execute the operations and calculate the so of each node in the dependency tree of the sentence, we start by initializing the so of each word using a subjective lexicon, in the manner of traditional unsupervised approaches (Turney, 2002).

Then, we traverse the parse tree in postorder, applying Algorithm 1 to update semantic orientations when visiting each node i. In this algorithm, O is the set of compositional operations defined in our system, A_i is a priority queue of the compositional operations to be applied at node i (because i is their destination node); and Q_i is another priority queue of compositional operations to be queued for upper levels at node i (as i is not yet their destination node). Push inserts o in a priority queue and pop pulls the operation with the highest priority (ties are broken by giving preference to the operation that was queued earlier). When visiting a node, a push into Q_i (Algorithm 1, line 7) is executed when the node i triggers an operation o that must be executed at the ancestor of i located δ levels upward from it. A push into A_i (Algorithm 1, line 9) is executed when the node i triggers an operation that must be executed at that same node i (i. e. $\delta = 0$). On the other hand, at node i, the algorithm must also decide what to do with the operations coming from $\text{children}_T(i)$. Thus, a push into A_i (Algorithm 1, line 13) is made when an operation from a child has reached its destination node (i. e. $\delta - 1 = 0$), so that it must be applied at this level. A push into Q_i (Algorithm 1, line 15) is made when the operation has still not reached its destination node and must be spread $\delta - 1$ more levels up.

5.2 FROM THEORY TO PRACTICE

At a practical level, the set of compositional operations are specified using a simple XML file:

Algorithm 1 Compute SO of a node

procedure COMPUTE(i, O ,T)
▷ Initialization of queues
  $A_i \leftarrow []$
  $Q_i \leftarrow []$
▷ Enqueue operations triggered by node i:
  for $o = (\tau, C, \delta, \pi, S)$ in O **do**
    if $C(i)$ **then**
      if $\delta > 0$ **then**
        $push((\tau, C, \delta, \pi, S), Q_i)$
      else
        $push((\tau, C, \delta, \pi, S), A_i)$
▷ Enqueue operations coming from child nodes:
  for c in $children_T(i)$ **do**
    for $o = (\tau, C, \delta, \pi, S)$ in Q_c **do**
      if $\delta - 1 = 0$ **then**
        $push((\tau, C, \delta - 1, \pi, S), A_i)$
      else
        $push(\tau, C, \delta - 1, \pi, S), Q_i)$
▷ Execute operations that have reached their destination node:
  while A_i is not empty **do**
    $o = (\tau, C, \delta, \pi, S) \leftarrow pop(A_i)$
    for j in $S(i)$ **do**
      $\sigma_j \leftarrow \tau(\sigma_j)$
▷ Join the SOs for node *i* and its children:
  $\sigma_i \leftarrow \sigma_i + \sum_{c \in children_T(i)} \sigma_c$

- `<forms>`: Indicates the tokens to be taken into account for the condition C that triggers the operation. Regular expressions are supported.
- `<dependency>`: Indicates the dependency types taken into account for C.
- `<postags>`: Indicates the PoS tags that must match to trigger the rule.
- `<rule>`: Defines the operation to be executed when the rule is triggered.
- `<levelsup>`: Defines the number of levels from i to spread before applying o.
- `<priority>`: Defines the priority of o when more than one operation needs to be applied over i (a larger number implies a bigger priority).

In addition, we need to integrate or train existent tools in order to create an end-to-end universal unsupervised software. We proceed to review them in §5.2.1.

5.2.1 *NLP tools for universal unsupervised sentiment analysis*

The following resources can serve us as the starting point to carry out state-of-the-art universal, unsupervised and syntactic sentiment analysis. [3]

Lexical resources

With respect to multilingual subjectivity lexica, there are a number of alternatives: SentiStrength (subjective data for up to 34 languages); the Chen and Skiena (2014) approach, which introduced a method for building sentiment lexicons for 136 languages; or SentiWordNet (Esuli and Sebastiani, 2015), where each synset from WordNet is assigned an objective, positive and negative score. Our implementation supports the lexicon format of SentiStrength, which can be plugged directly into the system. Additionally, we provide the option to create different dictionary entries depending on PoS tags to avoid conflicts between homonymous words (e.g. *'I'm fine'* versus *'They gave me a fine'*).

Tokenization

The system developed by Gimpel et al. (2011) is used for tokenizing. Although initially intended for English tweets, we have observed that it also performs robustly for many other language families (Romance, Slavic, etc.).

PoS tagging and dependency parsing

We will be using the monolingual parsers presented in §2.4 (Table 3) to obtain the syntactic structure of the documents evaluated in the experiments section. With respect to the PoS-taggers, we follow a similar approach: we use the universal tagset of the CPOSTAG column of the Universal Treebanks v2.0 (McDonald et al., 2013) and train a model relying on the Toutanova and Manning (2000) maximum-entropy tagger.

5.2.2 *Practical compositional operations*

We presented above a formalism to define arbitrarily complex compositional operations for unsupervised SA over a dependency tree. In this section, we show the definition of the most important rules that we used to evaluate our system. In practical terms, this implies studying how syntactic constructions that modify the sentiment of an expression are represented in the annotation formalism used for the

3 This is not an exhaustive list of available resources nor it plans to be. It just illustrates some of the most well-known options freely available for the research community.

training of the dependency parser, in this case, Universal Treebanks. We are using examples following those universal guidelines, since they are available for more than 40 languages and, as shown in § 5.3, the same rules can be competitive across different languages.

Intensification

As explained in Chapter 4, intensification amplifies or diminishes the sentiment of a word or phrase. Traditional lexicon-based methods handle most of these cases with simple heuristics, which might turn into misclassifications due to ambiguous cases. In said chapter we addressed the problem for Spanish, but we were limited to this language and the Ancora dependency structure.

Universal compositional operations overcome this problem in a multilingual setting without the need of any heuristic. A dependency tree already shows the behavior of a word within a sentence thanks to its dependency type, and it shows the role of a word independently of the language. Figure 6 shows graphically how universal dependencies represent the cases discussed above these lines. Formally, the operation for these forms of intensification is: $(\mathit{weighting}_{\beta}, w \in \text{intensifiers} \wedge t \in \{\text{ADV,ADJ}\} \wedge d \in \{\text{advmod,amod,nmod}\}, 1, 3, \mathit{dest} \cup \mathit{lm\text{-}branch}^{\text{acomp}})$, with the value of β depending on the strength of the intensifier as given by the sentiment lexicon.

'But' clauses

Compositional operations can also be defined to manage more challenging cases, such as clauses introduced by *'but'*, considered as a special case of intensification by different authors (Brooke, Tofiloski, and Taboada, 2009). It is assumed that the main clause connected by *'but'* becomes less relevant for the reader (e.g. *'It is expensive,* ***but*** *I love it'*). Figure 7 shows our proposed composition operation for this clause, formally: $(\mathit{weighting}_{\beta}, w \in \{\text{but}\} \wedge t \in \{\text{CONJ}\} \wedge d \in \{\text{cc}\}, 1, 1, \mathit{subjl})$ with $\beta = -0.25$. Note that the priority of this operation ($\pi = 1$) is lower than that of intensification ($\pi = 3$), since we first need to process intensifiers, which are local phenomena, before resolving adversatives, which have a larger scope.

Negation

As introduced in Chapter 4, dependency types help us to determine which nodes should act as negation and which should be its scope of influence. For brevity, we only illustrate some relevant negation cases and instructional examples that follow the universal treebank structure in Figure 8. Formally, the proposed compositional operation to tackle most forms of negation under universal guidelines is: $(\mathit{shift}_{\alpha}, w \in \text{negations} \wedge t \in U \wedge d \in \{\text{neg}\}, 1, 2, \mathit{dest} \cup \mathit{lm\text{-}branch}^{\text{attr}} \cup \mathit{lm\text{-}branch}^{\text{acomp}} \cup \mathit{subjr})$, where U represents the universal tag set.

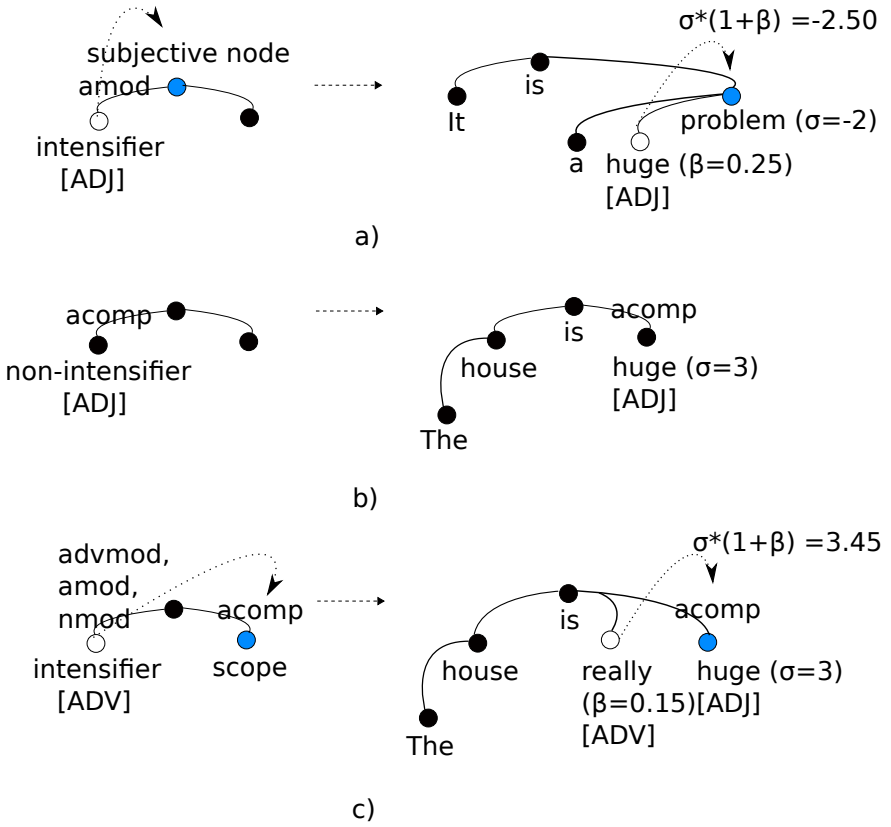


Figure 6: Skeleton for intensification compositional operations (2.a, 2.c) and one case without intensification (2.b), together with examples annotated with universal dependencies. Semantic orientation values are for instructional purposes only. In 2.a, *'huge'* is a term considered in a list of intensifiers, labeled as an ADJ, whose dependency type is *amod*, matching the definition of the intensification compositional operation. As a result, the o for intensification is triggered, spreading $\delta = 1$ levels up (i. e. up to *'problem'*) and amplifying the σ of *dest* (the first scope of the operation that matches, i. e. *'problem'*) by $(1+\beta)$. In 2.b, *'huge'* is again a word occurring in the intensifier list and tagged as an ADJ, but its dependency type is *acomp*, which is not considered among the intensification dependency types. As a result, no operation is triggered and the word is treated as a regular word (introducing its own so rather than modifying others). In 2.c, *'really'* is the term acting as intensifier, triggering again an intensification operation on the node $\delta = 1$ levels up from it (*'is'* node). Differently from 2.a, in this case the scope dest is not applicable since the word *'is'* is not subjective, but there is a matching for the second candidate scope, the branch labeled as acomp (the branch rooted at *'huge'*), so the σ associated with that node of the tree is amplified.

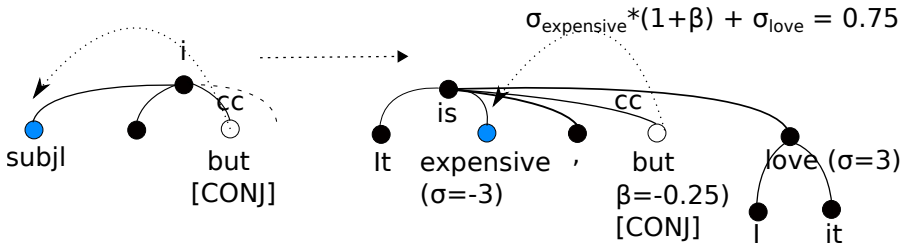


Figure 7: Skeleton for *'but'* compositional operation illustrated with one example according to universal dependencies. The term *'but'* matches the word form, tag and dependency types required to act as a sentence intensifier, so the compositional operation is queued to be applied $\delta = 1$ levels upward (i. e. at the *'is'* node). The scope of the operation is the first subjective branch that is a left child of said *'is'* node (i. e. the branch rooted at *'expensive'*). As a result, the σ rooted at this branch is diminished by multiplying it by $(1+\beta)$ (note that β is negative in this case) and the resulting value is added to the σ computed at *'is'* for the rest of the subjective children.

The priority of negation ($\pi = 2$) is between those of intensification and *'but'* clauses because its scope can be non-local, but it does not go beyond an adversative conjuction.

Irrealis

Irrealis denotes linguistic phenomena used to refer to non-factual actions, such as conditional, subjunctive or desiderative sentences (e.g. *'He **would** have died **if** he hadn't gone to the doctor'*). It is a very complex phenomenon to deal with, and systems are either usually unable to tackle this issue or simply define rules to ignore sentences containing a list of irrealis stop-words (Taboada et al., 2011). We do not address this phenomenon in detail in this study, but only propose a rule to deal with *'if'* constructions (e.g. *'if I die [...]'* or *'if you are happy'*, considering that the phrase that contains it should be ignored from the final computation. Formally: $(weighting_{\beta}, w \in \{\text{if}\} \wedge \text{t} \in \text{U} \wedge \text{d} \in \{\text{mark}\}, 2, 3, dest \cup subjr)$. Its graphical representation would be very similar to intensification (see Figures 5 a) and e)).

5.2.3 *Practical computation*

Example 17 represents an analysis of our introductory sentence *'He is not very handsome, but he has something that I really like'*, showing how compositional operations accurately capture semantic composition. [4]

EXAMPLE 17 (Example of a semantic orientation analysis of a sentence applying universal unsupervised prediction by UUUSA). For the sake

4 The UUUSA system released together with this thesis shows an equivalent ASCII text representation that can be obtained on the command line.

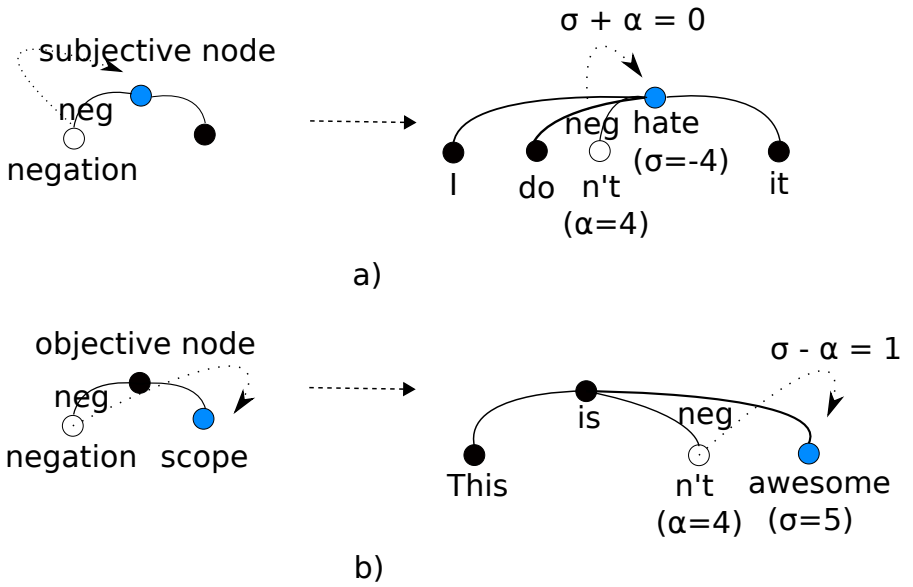
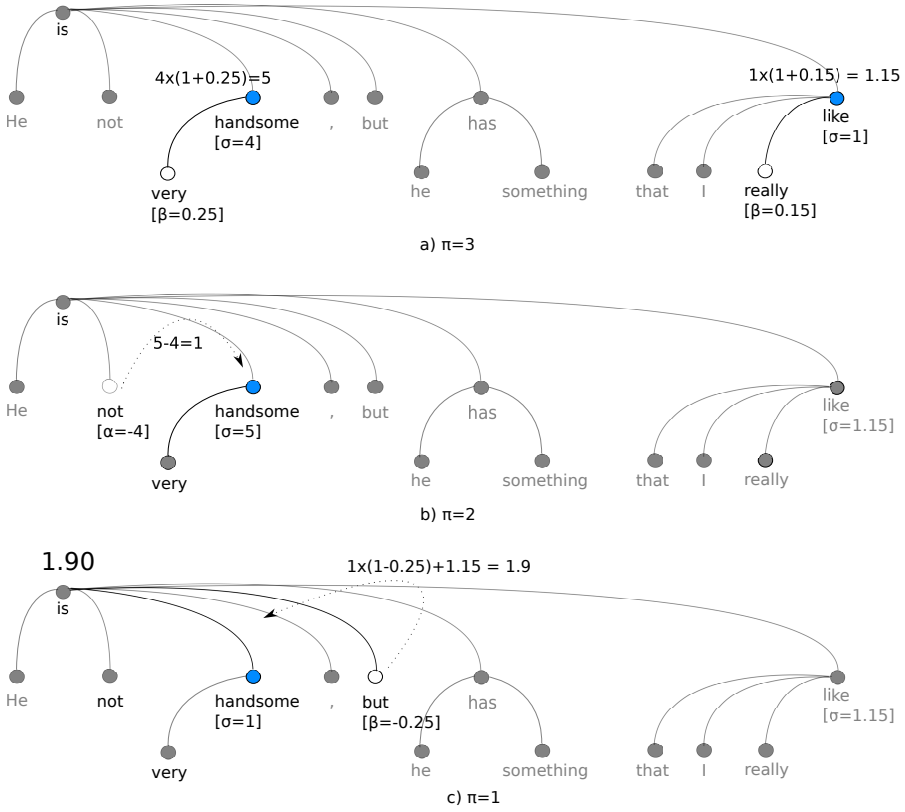


Figure 8: Skeleton for negation compositional operations illustrated together with one example. In 5.a, the term *'n't'* matches the word form of a negator and its dependency type is *neg*, queuing a negation compositional operation to be applied $\delta = 1$ levels upward (i. e. at the *'hate'* node). The first candidate scope for that operation matches, because dest is a subjective word (*'hate'*), shifting the σ of such word according to the definition of our $\texttt{shift}_{\alpha}(\sigma)$ transformation function. In a similar way, in 5.b, *'n't'* also acts a negator term, but in this case the candidate scope that matches is the second one (i. e. *lm-branch*$^{\text{attr}}$).

of clarity, the real post-order traversal is not illustrated. Instead we show an (in this case) equivalent computation by applying all operations with a given priority, π, at the same time, irrespective of the node. Semantic orientation, intensification and negation values are extracted from the dictionaries of Taboada et al. (2011).



Phase a) shows how the intensification is computed on the branches rooted at *'handsome'* and *'like'*. Phase b) shows how the negation shifts the semantic orientation of the attribute (again, the branch rooted at *'handsome'*). Phase c) illustrates how the clause *'but'* diminishes the semantic orientation of the main sentence, in particular the semantic orientation of the attribute, the first left subjective branch of its head. Elements that are not playing a role in a specific phase appear dimmed. One of the interesting points in this example comes from illustrating how three different phenomena involving the same branch (the attribute *'handsome'*) are addressed properly thanks to the assigned π.

Additionally, we below show the internal state and so updates made by the Algorithm 1:

Step	**Word**$_{index}$	**A**$_{word(\delta,\pi)}$	**Q**$_{word(\delta,\pi)}$	σ_{word}	$\sigma_{word} \leftarrow A$
1	He$_1$	[ ]	[ ]	0	0
2	not$_3$	[ ]	[$N_{not(1,2)}$]	0	0
3	very$_4$	[ ]	[$I_{very(1,3)}$]	0	0
4	handsome$_5$	[$I_{very(0,3)}$]	[ ]	4	5
5	,$_6$	[ ]	[ ]	0	0
5	but$_7$	[ ]	[$I_{but(1,1)}$]	0	0
6	he$_8$	[ ]	[ ]	0	0
7	something$_{10}$	[ ]	[ ]	0	0
8	has$_9$	[ ]	[ ]	0	0
9	I$_{12}$	[ ]	[ ]	0	0
10	that$_{11}$	[ ]	[ ]	0	0
11	really$_{12}$	[ ]	[$I_{really(1,3)}$]	0	0
12	like$_{13}$	[$I_{really(1,3)}$]	[ ]	1	1.15
13	is$_2$	[$N_{not(0,2)}$, $I_{but(0,1)}$]	[ ]	0	1.90

Each row corresponds to a step in which a node (Word$_{index}$) is visited in the postorder traversal. Columns $A_{word(\delta,\pi)}$ and $Q_{word(\delta,\pi)}$ show the state of the queues after the enqueuing operations, but before A is emptied (i. e. immediately before line 16 of Algorithm 1). The σ_{word} column shows the so of the visited node at that same point in time, and $\sigma_{word} \leftarrow A$ is the new so that is assigned by applying compositional operations and joining the sos of children (lines 16-20 of Algorithm 1). N and I refer to negation and intensification operations.

□

It is hard to measure the coverage of our rules and the potential of these universal compositional operations, since it is possible to define arbitrarily complex operations for as many relevant linguistic phenomena as wished. In this line, Poria et al. (2014) define a set of English sentic patterns to determine how sentiment flows from concept to concept in a variety of situations (e.g. relations of complementation, direct nominal objects, relative clauses, ...) over a dependency tree following the De Marneffe and Manning (2008) guidelines. The main difference of our work with respect to Poria et al. (2014) is that they present predefined sets of linguistic patterns for language-specific SA, whereas our approach is a theoretical formalism to define arbitrarily complex patterns given tagging and parsing guidelines, which has been implemented and tested on a universal set of syntactic annotation guidelines that work across different languages (see §5.3).

Under this approach, switching the system from one language to another only requires having a tagger and a parser following the Universal Treebanks (v2.0) guidelines and a subjectivity lexicon, but compositional operations remain unchanged (as shown in §5).

The performance of the algorithm might vary according to the quality of the resources on which it relies. Mistakes committed by the tagger and the parser might have some influence on the approach.

However, preliminary experiments on English texts show that having a parser with a LAS over 75% is enough to properly exploit compositional operations. With respect to the lexicalized parsing (and tagging) models, usually a different model is needed per language, even when using universal guidelines. In this respect, recent studies (Ammar et al., 2016; Guo et al., 2016; Vilares, Gómez-Rodríguez, and Alonso, 2016) have showed how it is possible to train a single model on universal treebanks to parse different languages with state-of-the-art results. This makes it possible to universalize one of the most relevant previous steps of our approach. The same steps can be taken to train multilingual tagging models (Vilares, Gómez-Rodríguez, and Alonso, 2016).

Adapting or creating new compositional operations for other tagging and parsing guidelines different from Universal Treebanks only requires: (1) becoming familiar with the new tag and dependency sets to determine which tags and dependency types should be included in each C, and (2) manually inspecting sentences parsed with the target guidelines to detect if they give a different structural representation of relevant phenomena. In this case, a new set of S, π or δ values may be needed, so that we can correctly traverse the tree and determine scopes on such dependency structure. At the moment, new practical operations need to be added manually, by defining them in the XML file.

5.3 EXPERIMENTS SHARING COMPOSITIONAL OPERATIONS

We compare our algorithm with respect to existing approaches on three languages: English, Spanish and German.

The availability of corpora and other unsupervised SA systems for English and Spanish enables us to perform a richer comparison than in the case of German, where we only have an *ad-hoc* corpus.

We compare our algorithm with respect to two of the most popular and widely used unsupervised systems: (1) SO-CAL Taboada et al. (2011), a language-dependent system available for English and Spanish guided by lexical rules at the morphological level, and (2) SentiStrength, a multilingual system that does not apply any PoS tagging or parsing step in order to be able to do multilingual analysis, relying instead on a set of subjectivity lexica, snippet-based rules and treatment of non-grammatical phenomena (e.g. character replication). Additionally, for the Spanish evaluation, we also took into account the system developed in Chapter 4, SSSA .

For comparison against state-of-the-art supervised approaches, we consider the deep recursive neural network presented by Socher et al. (2013), trained on a movie sentiment treebank (English). To the best of our knowledge, there are no semantic compositional supervised methods for Spanish and German.

Accuracy is used as the evaluation metric for two reasons: (1) it is adequate for measuring the performance of classifiers when the chosen corpora are balanced and (2) the selected systems for comparison also report their results using this metric.

5.3.1 *Datasets*

We selected the following standard English corpora for evaluation:

- Taboada and Grieve (2004) corpus. The same SFU Spanish review corpus used in Chapter 4. A general-domain collection of 400 long reviews (200 positive, 200 negative) about hotels, movies, computers or music among other topics, extracted from epinions.com.
- Pang and Lee (2004) corpus: A corpus of 2 000 long movie reviews (1 000 positive, 1 000 negative).
- Pang and Lee (2005) corpus: A corpus of short movie reviews (sentences). In particular, we used the test split used by Socher et al. (2013), removing the neutral ones, as they did, for the binary classification task (total: 1 821 subjective sentences).

To show the universal capabilities of our system we include an evaluation for Spanish using the corpus presented by Brooke, Tofiloski, and Taboada (2009) (200 positive and 200 negative long reviews from ciao.es). For German, we rely on a dataset of 2 000 reviews (1 000 positive and 1 000 negative reviews) extracted from Amazon.

As subjectivity lexica, we use the same dictionaries used by SO-CAL for both English (2 252 adjectives, 1 142 nouns, 903 verbs, 745 adverbs and 177 intensifiers) and Spanish (2 049 adjectives, 1 333 nouns, 739 verbs, 594 adverbs and 165 intensifiers). For German, we use the German SentiStrength dictionaries (Momtazi, 2012) instead (2 677 stems and 39 intensifiers), as Brooke, Tofiloski, and Taboada (2009) dictionaries are not available for languages other than Spanish or English. These are freely available resources that avoid the need to collect subjective words, intensifiers or negators. We just take those resources and directly plug them into our system. The weights were not tuned or changed in any way. To test the soundness of our theoretical formalism and the practical viability and competitiveness of its implementation, it does not matter what resource is chosen. We could have selected other available lexical resources such as SentiWordNet. The motivation for choosing SentiStrength (and SO-CAL) dictionaries is purely evaluative. We have compared our model with respect to other three state-of-the-art and widely used SA systems that use said resources. Our aim is not to evaluate our algorithm over a variety of different lexical resources, but to check if our universal system and compositional operations can compete with existing unsupervised systems under the same conditions (namely, using the same dictionaries and analogous sets of rules). The list of emoticons from Sen-

tistrength is also used as a lexical resource. If a term does not appear in these dictionaries, it will not have any impact on the computation of the so. [5] The content of these dictionaries and their parameters are not modified or tuned.

5.3.2 *Evaluation*

Comparison to unsupervised approaches

Table 16 compares the performance of our model with respect to SentiStrength[6] and so-cal on the Taboada and Grieve (2004) corpus. With respect to so-cal, results show that our handling of negation and intensification provides better results (outperforming so-cal by 3.25 percentage points overall). With respect to SentiStrength, our system achieves better performance on long reviews.

Table 17 compares these three unsupervised systems on the Pang and Lee (2004) corpus, showing the robustness of our approach across different domains. Our system again performs better than so-cal for negation and intensification (although it does not behave as well when dealing with irrealis, probably due to the need for more complex compositional operations to handle this phenomenon), and also better than SentiStrength on long movie reviews.

Rules	**SentiStrength**	**so-cal**	**uuusa**
Baseline	n/a	**65.50**	65.00
+negation	n/a	67.75	**71.75**
+intensification	66.00	69.25	**74.25**
+irrealis	n/a	71.00	**73.75**

Table 16: Accuracy on the Taboada and Grieve (2004) corpus (uuusa vs. other methods). We only provide one row for SentiStrength since we are using the standard configuration for English (which already includes negation and intensification functionalities).

Table 18 compares the performance of our universal approach on a different language (Spanish) with respect to: Spanish SentiStrength (Chapter 3), the Spanish so-cal Brooke, Tofiloski, and Taboada (2009) and sssa , a syntactic language-dependent system presented in Chapter 4. We used exactly the same set of compositional operations as used for English (only changing the list of word forms for negation, intensification and *'but'* clauses, as explained in §5.1.1). Our universal system again outperforms SentiStrength and so-cal in its Spanish

5 Out-of-vocabulary words are not given a special treatment at the moment.

6 We used the default configuration, which already applies many optimizations. We set the length of the snippet between a negator and its scope to 3, based on empirical evaluation, and applied the configuration to compute sentiment on long reviews.

Rules	**SentiStrength**	**so-cal**	**uuusa**
Baseline	N/A	**68.05**	67.77
+negation	N/A	70.10	**71.85**
+intensification	56.90	73.47	**74.00**
+irrealis	N/A	**74.95**	74.10

Table 17: Accuracy on the Pang and Lee (2004) test set (UUUSA vs. other methods)

Rules	**SentiStrength**	**so-cal**	**uuusa**	**sssa**
Baseline	N/A	N/A	**63.00**	61.80
+negation	N/A	N/A	**71.00**	N/A
+intensification	73.00	N/A	74.25	**75.75**
+irrealis	N/A	74.50	**75.75**	N/A

Table 18: Accuracy on the Spanish Brooke, Tofiloski, and Taboada (2009) test set with a variety of options enabled for various methods

version. The system also obtains results very similar to the ones reported in Chapter 4, even though their system is language-dependent and the set of rules is fixed and written specifically for Spanish.

In order to check the validity of our approach for languages other than English and Spanish, we have considered the case of German. It is worth noting that the authors of this work have no notions of German at all. In spite of this, we have been able to create a state-of-the-art unsupervised SA system by integrating an existing sentiment lexicon into the framework that we propose in this chapter.

We use the German SentiStrength system (Momtazi, 2012) for comparison. The use of the German SentiStrength dictionary, as mentioned in Section 5.3.1, allows us to show how our system is robust when using different lexica. Experimental results show an accuracy of 72.75% on the Amazon review dataset when all rules are included, while SentiStrength reports 69.95%. Again, adding first negation (72.05%) and then intensification (72.85%) as compositional operations produced relevant improvements over our baseline (69.85%). The results are comparable to those obtained for other languages, using a dataset of comparable size, reinforcing the robustness of our approach across different domains, languages, and base dictionaries.

Comparison to supervised approaches

Supervised systems are usually unbeatable on the test portion of the corpus with which they have been trained. However, in real applica-

tions, a sufficiently large training corpus matching the target texts in terms of genre, style, length, etc. is often not available; and the performance of supervised systems has proven controversial on domain transfer applications (Aue and Gamon, 2005).

Table 19 compares our universal unsupervised system to Socher et al. (2013) on a number of corpora: (1) the collection used in the evaluation of the Socher et al. system (Pang and Lee, 2005), (2) a corpus of the same domain, i. e. movies (Pang and Lee, 2004), and (3) the Taboada and Grieve (2004) collection. Socher et al.'s system provides sentence-level polarity classification with five possible outputs: *very positive*, *positive*, *neutral*, *negative*, *very negative*. Since the Pang and Lee (2004) and Taboada and Grieve (2004) corpora are collections of long reviews, we needed to collect the global sentiment of the text. For the document-level corpora, we count the number of outputs of each class [7] (*very positive* and *very negative* count double, *positive* and *negative* count one and *neutral* counts zero). We take the majority class, and in the case of a tie, it is classified as negative.[8]

The experimental results show that our approach obtains better results on corpora (2) and (3). It is worth mentioning that our unsupervised compositional approach outperformed the supervised model not only on an out-of-domain corpus, but also on another dataset of the same domain (movies) as the one where the neural network was trained and evaluated. This reinforces the usefulness of an unsupervised approach for applications that need to analyze a number of texts coming from different domains, styles or dates, but there is a lack of labeled data to train supervised classifiers for all of them. As expected, Socher et al. (2013) is unbeatable for an unsupervised approach on the test set of the corpus where it was trained. However, our unsupervised algorithm also performs very robustly on this dataset.

5.4 EXPERIMENTS SHARING LEXICA, PARSING MODELS AND COMPOSTIONAL OPERATIONS

In the previous section, we have shown how the same set of compositional operations can be shared across different languages. In this section, we go one step beyond and show how it is possible to create an effective end-to-end multilingual SA analysis system, where in addition to the compositional operations, also the subjectivity lexica,

7 When trying to analyze the document-level corpora with Socher et al.'s system, we had *out-of-memory problems* on a 64-bit Ubuntu server with 128GB of RAM memory, so we decided to choose a counting approach instead over the sentences of such corpora.

8 These criteria were selected empirically. Assigning the positive class in the case of a tie was also tested, as well as not doubling the *very positive* and *very negative* output, but these settings produced similar or worse results with the Socher et al. (2013) system.

Corpora	Socher et al. (2013)	uusa
Origin corpus of Socher et al. (2013) model		
Pang and Lee 2005 Pang and Lee (2005)	**85.40**	75.07
Other corpora		
Taboada and Grieve (2004)	62.00	**73.75**
Pang and Lee 2004 Pang and Lee (2004)	63.80	**74.10**

Table 19: Accuracy on different corpora for Socher et al. (2013) and UUUSA . On the Pang and Lee 2005 Pang and Lee (2005) collection, our detailed results taking into account different compositional operations were: 73.75 (baseline), 74.13 (+negation), 74.68 (+intensification) and 75.07 (+irrealis)

and the parsing and tagging models can be shared effectively. In particular, we present a model, called SISA, that analyzes five official languages in the Iberian Peninsula: Basque (*eu*), Catalan (*ca*), Galician (*gl*), Portuguese (*pt*) and Spanish (*es*). We proceed by combining existing subjectivity lexica, training a *single Iberian* tagger and parser, and defining a set of Iberian syntax-based rules. As a result, we are obtaining:

1. A single set of syntactic compositional operations to handle linguistic phenomena across five Iberian languages.
2. The first end-to-end multilingual syntax-based SA system that analyzes five official languages of the Iberian Peninsula. This is also the first evaluation for SA that provides results for some of them.

In the context of the Iberian Peninsula, much of the literature has focused on Spanish (Brooke, Tofiloski, and Taboada, 2009; Gamallo, García, and Fernández Lanza, 2013; Hurtado, Pla, and Buscaldi, 2015; Saralegi and San Vicente, 2013). Portuguese has also attracted interest, focusing on political domains (Silva et al., 2009), development of resources (Balage Filho, Pardo, and Aluísio, 2013; Souza et al., 2011) and exploring the influence of NLP in SA (Souza and Vieira, 2012). For Basque and Catalan literature is scarce and limited to the development of resources (Bosco et al., 2016; Cruz et al., 2014a; San Vicente and Saralegi, 2016). For Galician, we present the very first insights.

We below present how to build SISA, from the bottom (subjectivity lexica, tagging and dependency parsing) to the top levels (application of compositional operations to compute the final SO), and also the datasets we are using for the evaluation.

5.4.1 *Datasets*

The following corpora will be used to evaluated SISA.

• *The* SFU *Spanish review corpus* (Brooke, Tofiloski, and Taboada, 2009): A set of 400 long reviews (200 positive, 200 negative) from different domains such as movies, music, computers or washing machines.

• *Portuguese SentiCorpus-PT 0.1* (Carvalho et al., 2011): A collection of comments from the Portuguese newspaper *Público* with polarity annotation at the entity level. As our system assigns the polarity at the sentence level, we selected the SentiCorpus sentences with (a) only one SO and (b) with > 1 SO iff all of them were the same, generating a corpus with 2 086 (from 2 604) sentences.

• *Basque opinion dataset* (San Vicente and Saralegi, 2016): Two small corpora in Basque containing news articles and reviews (music and movie domains). We merged them to create a larger dataset, containing a total of 224 reviews.

In addition, due to the lack of available sentence- or document-level corpora for Catalan or Galician, we opted for synthetic corpora:

• *Synthetic Catalan* SFU : An automatically translated version to *ca* of the Spanish SFU , with 5% of the words from the original corpus considered as unknown by the translation tool.

• *Synthetic Galician* SFU : An automatically translated version to *gl* of the Spanish SFU ($\approx$ 6.4% of the words not translated).

5.4.2 *Multilingual subjectivity lexica*

SISA needs multilingual polarity lexica in order to predict the sentiment of a text. We used two sets of monolingual lexica as our starting points:

1. Brooke, Tofiloski, and Taboada (2009) dictionaries: It contains SO's for subjective words that range from 1 to 5 for positive and negative terms. We translated it to *ca*, *eu*, *gl* and *pt* using *apertium* (Forcada et al., 2011). We removed the unknown words and obtained the numbers in Table 20.[9]

2. Cruz et al. (2014a) lexicon: Multi-layered lexica (not available for *pt*) with SO's where each layer contains a larger number of terms, but less trustable. We used the seventh layer for each language. As *eu*, *ca* and *gl* files have the same PoS-tag for adverbs and adjectives, they were automatically classified using monolingual tools (Agerri, Bermudez, and Rigau, 2014; Garcia and Gamallo, 2015; Padró and Stanilovsky, 2012) (Table 21 contains the statistics). SO's (originally from 0 to 1) were linearly transformed to the scale of the Brooke, Tofiloski, and Taboada (2009) dictionaries.

The Brooke, Tofiloski, and Taboada (2009) and Cruz et al. (2014a) lexica for each language were combined to obtain larger monolingual

9 We used the original *apertium* outputs, except for the *pt* and *gl* lexica (manually reviewed by a linguist).

Tag	es	pt	ca	eu	gl
ADJ	2 045	1 865	1 686	1 757	2 002
NOUN	1 323	1 183	1 168	1 211	1 270
ADV	594	570	533	535	599
VERB	739	688	689	563	723

Table 20: Size of the Brooke, Tofiloski, and Taboada (2009) (single words) lexica after being translated

Tag	es	ca	eu	gl
ADJ	2 558	1 619	22	1 530
NOUN	2 094	1 535	1 365	579
ADV	117	23	3	26
VERB	603	500	272	144

Table 21: Size of the resulting Cruz et al. (2014a) lexica after processing.

resources, and these were in turn combined into a common Iberian lexicon (see Table 22). When merging lexica, we must consider that:

Tag	es	pt	ca	eu	gl	Iberian
ADJ	3 775	1 865	2 704	1 529	2 990	9 385
NOUN	3 079	1 183	2 377	2 392	1 684	8 733
ADV	665	570	545	485	612	1 891
VERB	1 177	688	1 034	728	801	2 998

Table 22: Size of the final lexica used by SISA.

1. In monolingual mergings, the same word can have different so's. For example, the Catalan adjective *'abandonat'* (*'abandoned'*) has -1.875 and -3 in Cruz et al. (2014a) and Brooke, Tofiloski, and Taboada (2009), respectively.

2. When combining lexica of different languages, the same word form might have different meanings (and so's) in each language. Merging them in a multilingual resource could be problematic. For example, the adjective *'espantoso'* has a value of -4.1075 in the combined *es* lexicon (*'frightening'*), and of -3.125 in the *gl* one (*'frightening'*), while the same word in the *pt* data (*'astonishing'*) has a positive value of 5. Note, however, that even if they could be considered very similar from a lexical or morphological perspective, many false friends have different spellings in each language, such as the negative *'vessar'* (*'to spill'*) in *ca* and the positive *'besar'* (*'to kiss'*) in *es*, so these cases end up not being a frequent problem (only 0.36% of the words have both positive and negative polarity in the monolingual lexica).

These two problems were tackled by averaging the polarities of words with the same form. Thus, the first monolingual mergings produced a balanced SO (e.g., *'abandonat'* has −2.4375 in the combined *ca* lexicon), while in the subsequent multilingual fusion, contradictory false friends have a final value close to *no polarity* (e.g., *'espantoso'*, with a SO of −0.7 in the Iberian lexicon). The impact of these mergings is analyzed in §5.4.4.

5.4.3 *Multilingual PoS tagging and dependency parsing*

For the compositional operations to be triggered, we first need to do the tagging and the dependency parse for a sentence. To do so, we trained an Iberian PoS-tagger and parser, i.e. single modules that can analyze Iberian languages without applying any language identification tool. Such multilingual taggers and parsers can be effectively trained following approach we introduced in §2.4. We are relying on the Universal Dependencies (Nivre et al., 2016) to train these tools, since they provide corpora for all languages studied in this section.

For the Iberian parser we followed the same methodology described in §2.4, whose performance (LAS/UAS) on the same UD test sets was: *pt* (78.78/84.50), *es* (80.20/85.23), *cat* (84.01/88.08), *eu* (62.01/71.64) and *gl* (75.65/82.11). The parsing results for Basque (with a high proportion of non-projective trees) were worse than expected. However, the parser trained based on our method selected a projective algorithm for training, as the average prevalence of non-projectivity across our five Iberian languages is low. We hypothesize that this is the main reason of the lower performance for this language. For the Iberian tagger we obtained the following accuracies (%) in the monolingual UD test sets using a similar angle (and training on the Toutanova and Manning (2000) tagger): *pt* (95.96), *es* (94.37), *ca* (97.41), *eu* (93.88) and *gl* (94.09).

We consider the compositional operations of intensification, subordinate adversative clauses, negation and *'if'* irrealis, the same practical set of compositional operations defined in §5.2.2, but adapted to the UD guidelines.

5.4.4 *Evaluation*

This section presents the results of the experiments we carried out with our system using both the monolingual and the multilingual lexica, compared to the performance of a supervised classifier for three of the five analyzed languages.

We performed different experiments on binary polarity classification for knowing (a) the accuracy of the system, (b) the impact of the merged resources, and (c) the impact of the universal operations in monolingual and multilingual settings:

1. SL-O: Single lexica, no operations (baseline).
2. ML-O: Multilingual lexica, no operations.
3. SL+O: Single lexica with universal operations.
4. ML+O: Multilingual lexica with universal operations.

The performance of our system was compared to LinguaKit (*LKit*), an open-source toolkit which performs supervised sentiment analysis in several languages (Gamallo, García, and Fernández Lanza, 2013).[10]

Lg	sl-o	sl+o	ml-o	ml+o	LKit
ES	60.00	75.75	63.75	**76.50**	58.75
CA	54.00	57.50	58.25	**73.00**	—
GL	60.75	**73.00**	60.00	70.00	50.25
EU	62.95	69.20	65.63	**72.32**	—
PT	60.50	**67.35**	57.29	65.01	60.55

Table 23: Results of SISA and different configurations vs *LKit* on different test sets of Iberian languages. In *LKit* we only evaluated the positive and negative results (it also classifies sentences with no polarity).

Table 23 contains the results of each of these models on the different corpora. The baseline (SL-O) obtained values between 54% (*ca*) and 62.95% (*eu*), which are competitive results when compared to those obtained by the supervised model.[11] As we are not aware of available SA tools for *ca*, we could not compare our results with other systems. For Basque, San Vicente and Saralegi (2016) evaluated several lexica (both automatically translated and extracted, as well as with human annotation) in the same dataset used in this section. They used a simple average polarity ratio classifier, which is similar to our baseline. Even if the lexica are different, their results are very similar to our SL-O system (63% *vs* 62,95%), and they also show that manually reviewing the lexica can boost the accuracy by up to 13%.

The central columns of Table 23 show the results of using universal operations and a merged lexicon in the same datasets. In *gl* and *pt* the best values were obtained using individual lexica together with syntactic operations, while the Iberian system achieved the best results in the other languages.

Table 24 summarizes the impact that the operations have in both the monolingual and the multilingual setting, as well as the differences in performance due to the fusion process. Concerning the operations (columns 2 and 3), the results show that using the same set of universal operations improves the performance of the classifier in all the languages and settings. Their impact varies between 3.5 percentage points (*ca*) and more than 15 (*es*) and, for each language, the

10 https://github.com/citiususc/Linguakit

11 LinguaKit was intended for tweets (not long texts).

Lg	o(sl)	o(ml)	ml(-o)	ml(+o)
es	+15.75	+12.75	+3.75	+0.75
ca	+3.50	+14.75	+4.25	+15.5
gl	+12.25	+10.00	-0.75	-3.00
eu	+6.25	+6.69	+2.68	+3.12
pt	+6.85	+7.72	-3.21	-2.34

Table 24: Impact of the operations (o) with mono (SL) and multilingual lexica (ML) and of the ML with (+o) and without operations (-o)

operations provide a similar effect in monolingual and multilingual lexica (except for *ca*, with much higher values in the ML scenario).

The fusion of the different lexica had different results (columns 4 and 5 of Table 24): in *gl* and *pt*, it had a negative impact (between -0.75% and -3.21%) while in the other three the ML setting achieved better values (between 0.75 and 15.5 points, again with huge differences in *ca*). On average, using multilingual lexica had a positive impact of 1.3 (-o) and 2.8 points (+o).

As mentioned, *ca* has a different behaviour: the gain from operations when using monolingual lexica is about 3.50 points (lower than other languages), and the benefit of the ML lexicon without syntactic operations is of 4.25 points. However, when combining both the universal operations and the ML lexicon its performance increases ≈ 15 points, turning out that the combination of these two factors is decisive.

In sum, the results of the experiments indicate that syntactic operations defined by means of a harmonized annotation can be used in several languages with positive results. Furthermore, the merging of monolingual lexica (some of them automatically translated) can be applied to perform multilingual SA with little impact on performance when compared to language-dependent systems.

5.5 CONCLUSION

In this chapter, we have described, implemented and evaluated a novel model for universal and unsupervised sentiment analysis driven by a set of syntactic operations for semantic composition. Existing unsupervised approaches are purely lexical, their rules are heavily dependent on the language concerned or they do not consider any kind of natural language processing step in order to be able to handle different languages, using shallow rules instead.

To overcome these limitations, we introduce from a theoretical and practical point of view the concept of compositional operations, to define arbitrarily complex semantic relations between different nodes of a dependency tree. Universal part-of-speech tagging and dependency

parsing guidelines make it feasible to create multilingual sentiment analysis compositional operations that effectively address semantic composition over natural language sentences. The system is not restricted to any corpus or language, and by simply adapting or defining new operations it can be adapted to any other PoS tag or dependency annotation criteria.

We have compared our universal unsupervised model with state-of-the-art unsupervised and supervised approaches. Experimental results show: (1) that our algorithm outperforms two of the most commonly used unsupervised systems, (2) the universality of the model's compositional operations across different languages and (3) the usefulness of our approach on domain-transfer applications, especially with respect to supervised models.

In addition, we built a single symbolic syntactic system for polarity classification that analyzes five official languages of the Iberian peninsula. We show that with very little effort it is possible to obtain a competitive multilingual sentiment analysis system working on many languages.

Part III

MACHINE-LEARNING COMPOSITIONAL STRATEGIES

6

A SUPERVISED MODEL BASED ON GENERALIZATION FOR MONOLINGUAL SENTIMENT ANALYSIS

In Part ii, we introduced a number of knowledge-based models that considered lexical and syntactic information to determine the semantic orientation of texts. In this part, we follow a different perspective and consider a machine-learning approach to classify the polarity of tweets by using linguistic knowledge.

This chapter focuses on Spanish tweets although it could be equally applicable to any other language. The main contribution consists of building models which combine lexical, syntactic, psychometric and semantic knowledge to illustrate the performance that linguistic perspectives can achieve, ranging from shallow to deep knowledge. The system described in Chapter 4, initially intended for long reviews, is used to enrich the supervised models built in this chapter. We also introduce the concept of *enriched generalized dependency triplets*, a syntactic feature representation inspired on the Joshi and Penstein-Rosé (2009) generalized triples for identification of opinionated sentences on long reviews, that we use here for polarity classification tasks. We additionally explore how the size of the training set is relevant to properly exploit different linguistic features.

We also undertake a wide experimental evaluation, suggesting that a syntactic perspective outperforms pure lexicon-based methods if the training collection is large enough. Most of the results only focus on classifying tweets as positive, negative or objective, but we also provide some conclusions regarding a finer-grained classification.

A model implementing this approach can be found as a part of the miopia library[1].

6.1 DESCRIPTION

In this chapter we study how lexical and syntactic features can help improve polarity classification accuracy over Spanish tweets. In addition to the word forms, there exist several ways to extract complementary information to obtain better classifications. Many terms are associated with psychological properties, such as anxiety, anger or happiness. In the same line, morphological information can help discriminate between subjective and objective texts. For example, adjectives, adverbs or first person pronouns are a priori good indicators of opinionated texts. All this information is used and combined to

1 http://grupolys.org/software/MIOPIA/

create different supervised classifiers, in order to improve standard bag-of-terms approaches. Moreover, we hypothesize that by syntactically relating these kinds of information it is possible to capture more context, improving accuracy (Socher et al., 2013). For this purpose, we are using dependency parsing to identify relations between words in order to overcome the problem of many sentiment detection approaches, which take into account individual words, but not their context. To identify these relations we rely on a more relaxed concept of generalized dependency triplets.

6.1.1 *Generalized dependency triplets*

Our aim is to use dependency triplets to capture interesting patterns between terms, modeling common linguistic phenomena such as negation or intensification, and many others which are difficult to treat by symbolic and pure lexicon-based approaches. In general terms, figures of speech such as oxymoron are good examples of complex constructions that are uncommon, but should be taken into account by sentiment classifiers.

In Joshi and Penstein-Rosé (2009) the authors explore the effectiveness of dependency-based features on identifying opinionated sentences. They introduce the concept of *composite back-off features*, or *generalized dependency triplet features*; which is the term we are using to refer this method in this thesis: given a dependency triplet of the form $i \xrightarrow{d} j$, they propose generalizing either the head term (located at i) or the dependent term (located at j) to their respective part-of-speech tag. How the process works is shown in Example 18. Their approach obtains a statistically significant improvement when some of these generalized features are used in conjunction with word unigrams. Specifically, they obtained the best performance when applying generalization over the head term. They concluded generalizing the head is a better option because makes it possible to identify patterns in opinions about products, features or services. The dependency type does not play a role, in terms of generalization, in the work by Joshi and Penstein-Rosé (2009). In any case, keeping information about the dependency type which connects a pair of words could be useful, as a way to capture how people connect terms.

Example 18 (Example of the Joshi and Penstein-Rosé (2009) generalization method). In the sentences *'He is a smart boy'* and *'It's a smart television'*, Joshi and Penstein-Rosé (2009) approach generalizes the triplets $\mathtt{boy} \rightarrow \mathtt{smart}$ and $\mathtt{television} \rightarrow \mathtt{smart}$ to a single triplet of the form $\mathtt{noun} \rightarrow \mathtt{smart}$.

In this way, two triplets that have the same semantic meaning in terms of binary polarity classification are unified into one, while relations can still be captured.

□

6.1.2 *Generalized dependency triplets for sentiment analysis*

The idea of the concept of generalized composite features is presented by Joshi and Penstein-Rosé (2009). However, we consider this perspective is in itself incomplete for performing polarity classification on micro texts, for several reasons. Firstly, the authors worked on product reviews from Amazon, where vocabulary is more restricted and reduced than in other social media, and ungrammatical elements are not so frequent. In addition, they used their perspective on identifying opinionated sentences in that domain, but it was not intended nor evaluated for classifying sentiment, neither on long nor on micro-texts. In this respect, only generalizing to coarse PoS-tags can involve a loss of very useful information. In order to facilitate understanding, we will use examples in English to illustrate the relevant syntactic constructions in this and following sections, although the approach we are describing is designed for Spanish. An example where PoS-tagging generalization might cause significant losts of relevant information is shown below these lines.

EXAMPLE 19 (Example of losts of relevant information when applying Joshi and Penstein-Rosé (2009) generalization for polarity classification). Consider the sentence *'He makes a delicious villain'*. According to the method proposed by Joshi and Penstein-Rosé (2009), the triplet $villain \rightarrow delicious$ would be generalized as $noun \rightarrow delicious$ or $villain \rightarrow adjective$. However, this is not an optimal generalization. For example, selecting the option $villain \rightarrow adjective$ we are losing useful information because *'delicious'* provides sentiment by itself. However, if we try to use the original triplet, we will probably have sparsity problems because it is very unlikely that we have seen that specific combination of words and dependency relation in the training set. Finally, a base unigram approach would not be able to treat this sentence correctly, since the meaning of *'delicious villain'* can be different depending on whether these words appear together (which could be considered an oxymoron) or apart.

□

We adapt and enrich the initial concept of generalized dependency features, intended for detecting opinionated sentences, to improve the accuracy of lexicon-based sentiment classifiers. We incorporate various levels of generalization both for the head and the dependent term, instead of just using part-of-speech information. We also contemplate deleting the dependency type, keeping only the head and the dependent term, which could be considered as a syntactic n-gram. We express this formally in Definition 4.

DEFINITION 4. Let $i \xrightarrow{d} j$ be a regular dependency triplet, an **enriched generalized dependency triplet** is a triplet of the form $g(i, \zeta)\ del(\xrightarrow{d})\ g(j, \zeta)$, such that:

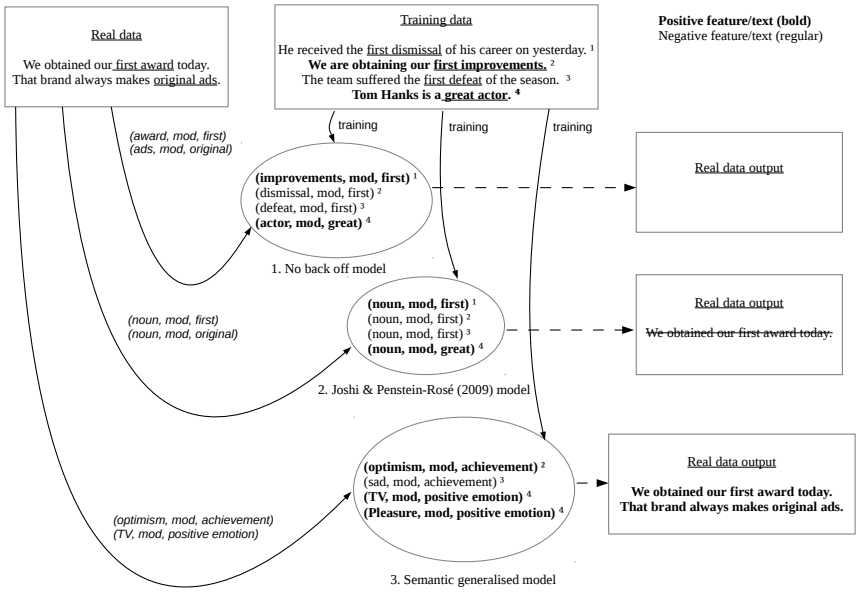
- $\zeta \in Z$ is the type of desired generalization for a word indexed at a position i, where $Z = \{word, lemma, postag, psychometric, blank\}$
- $g : W \times Z \rightarrow \{W, L, P, PS, \varnothing\}$ is a generalization function that transforms the word indexed at the position i to a value according to ζ, where W is the set of words, L the set of lemmas, P the set of possible PoS-tags, PS indicates a list of psychometric values and $\varnothing$ is the empty set when the generalization option is $blank$.
- $del : D \rightarrow \{D, \varnothing\}$ is a deletion function to decide if we keep or remove the dependency type as a part of the compositve feature, where D is the set of dependency types.

□

The goal here is to generalize composite features, but in such a way that we do not lose too much relevant semantic information. To the best of our knowledge, this is the first study on proposing this kind of composite generalized dependency triplets. Example 20 sketches some of the theoretical advantages of this approach.

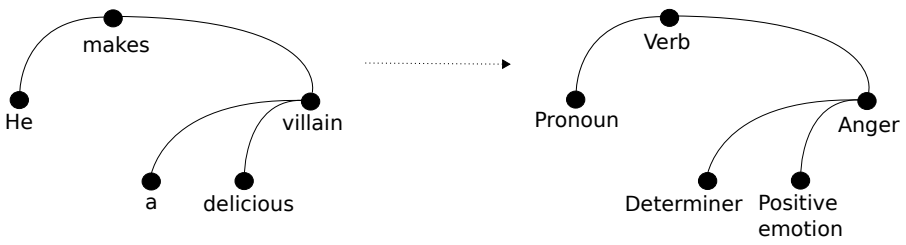
Example 20. The word *'villain'* could be assigned to the psychometric properties *negative emotion* or *anger*, and the term *'delicious'* could be classified as a *positive emotion*. Thus, in the sentence *'He makes a delicious villain'* we could extract the triplets $negative\ emotion \xrightarrow{mod} positive\ emotion$ and $anger \xrightarrow{mod} positive\ emotion$, which are purely semantic dependencies, but more generalizable than $villain \xrightarrow{mod} delicious$. Other examples of generalization options could be $negative\ emotion \rightarrow positive\ emotion$ and $anger \rightarrow positive\ emotion$ if we omit the dependency type, or $common\ noun \rightarrow positive\ emotion$ if we apply a different generalization for each term.

In the figure below, we display different dependency triplets generalized models illustrated with one example. We use both a hypothetical training dataset and a small test dataset.



The dependency type *mod* that appears in the dependency triplets is the short form of the syntactic function *modifier*. In the example, dependency triplets always contain this dependency type, because they are representing an adjective which modifies a noun. Our method generalizes the words *'ads'* and *'actor'* to the category *'TV'*, according to the avaliable resources. This is an example of one of a number of generalizations that would be made by the new method. In fact, generalizing *'actor'* to *'Films'* and *'ads'* to *'Newspaper'* or *'Magazine'* would be two additional and acceptable options. Underlined phrases refer to the triplets that are taken as features. Boldface text refers to either a positive text or a generalized dependency triplet which implies a favorable sentiment. The real data output box contains the expected results for each model, where the strikethrough text indicates that the prediction was wrong and omitted texts mean that the model was unable to assign any label, given the corresponding input.

And the graph below shows how our running example could be ideally represented based on psychometric properties, obtaining a more general graph that could match many other sentences with similar polarity:



□

6.1.3 *Classifier*

We are using a standard supervised classifier to train our model. In this chapter, we see the classifier as an abstraction, where given a set C of target classes and a set F of eligible features, a *bag-of-words* model will optimize $\arg\max_c h(c, f)$, where h is a hypothesis function, $c \in C$ and f an orderless set of features with $f \subseteq F$.

We have chosen an SMO classifier, an implementation of SVM proposed by Platt (1999) that is available in popular data mining software packages (e.g. Hall et al. (2009)). Preliminary experiments suggested that the SMO outperformed other implementations of SVM during the development process, and also other classification techniques such as Bayesian models or decision trees. In addition, we apply feature reduction. A lower number of features makes the training process faster and helps to avoid irrelevant attributes, which is especially important in noisy media such as Twitter. We relied on information gain (Mitchell, 1997) to decide the relevance of features in each model, selecting only those features with an information gain greater than zero.

6.1.4 *Features*

Selecting good discriminative features is crucial for a robust performance of linear classifiers and SVM's. We have defined several features in order to test the effectiveness of relating features by means of dependency parsing when they are used in conjunction with models based on lexical and semantic knowledge:

- *Sentiment information*: We use the information provided by SSSA (described in Chapter 4). In particular, the features the analyzer provides to the classifier for each tweet are: (1) its global semantic orientation; (2) the number of positive words that appear in the tweet; and (3) the number of negative words that appear in the tweet.
- *Words*: A widely used supervised approach is to consider tweets as a set of orderless words and to use them to feed a supervised classifier. Although simple, this strategy generally shows a robust performance.
- *Lemmas*: A natural extension of the previous approach is to first apply lemmatization which allows the number of features to be restricted. This can be useful in languages such as Spanish, where gender or number is expressed by declensions of nouns, adjectives or verbs. We rely on the collection of lemmas provided by the Ancora corpus (Taulé, Martí, and Recasens, 2008) to lemmatize words.
- *Lexical bi-grams*: In addition to uni-grams, we also performed experiments using bi-grams of words and lemmas.
- *Part-of-speech tags*: The use of PoS tags in polarity classification is a widely discussed issue in many studies (Pang, Lee, and Vaithyanathan,

2002; Spencer and Uchyigit, 2012). However, the utility of PoS tags by themselves is camouflaged because they are used in conjunction with other features (Pak and Paroubek, 2010; Zhang et al., 2009). We test the effectiveness of both fine and coarse part-of-speech tags.

- *Psychometric properties*: We introduce a perspective based on psychological knowledge. We rely on the dictionaries presented by Ramírez-Esparza et al. (2007). This lexicon distinguishes around 70 dimensions of human language. It provides information about psychometric properties of words (e. g. cognition mechanisms, anxiety, sexuality), but also considering topics (e. g. TV, family, religion) or even linguistic information (e. g. past, present and future tense, exclamations or questions). In this way, the verb *'imagine'* would represent a *cognition mechanism* and *insight*. This psychological linguistic resource is found in the LIWC software (Pennebaker, Francis, and Booth, 2001).
- *Dependency types*: We take only the identifiers of the dependencies occurring in the parse tree of each tweet. Thus we are not considering any information regarding the words linked by dependencies. In this case, we try to test if dependency types can be helpful by themselves to solve polarity classification tasks.
- *Syntactic triplets*: The models described above these lines will serve as a starting point from which to incorporate syntactic knowledge. Concretely, we represent syntactic information by means of generalized dependency-based features described in §6.1.2. The aim is to measure the effectiveness and sparsity problems of this type of features when they are used both separately and in conjunction with lexicon-based models. We test different levels of generalization over the head and the dependent word of a dependency triplet, including lemmas, psychometric properties and fine PoS-tags.

6.2 EXPERIMENTS

6.2.1 *Dataset*

The TASS general corpus is a collection of tweets which has been specifically annotated to perform polarity classification at a global level, presented at the Workshop on Sentiment Analysis at SEPLN (Villena-Román and García-Morera, 2013). It is a collection of Spanish tweets written by public figures, such as soccer players, politicians or journalists. Messages range from November 2011 to March 2012. The corpus is composed of a training set and a test set which contain 7 219 and 60 798 tweets, respectively. Each tweet is annotated with one of these six labels: strong positive (POS+), positive (POS), neutral (NEU), negative (NEG), strong negative (NEG+) and none (NONE). Neutral tweets refer to messages that contain both positive and negative ideas; whereas tweets labeled as NONE concern those that do not express any sentiment. The gold standard was generated by a pooling

of the submissions, followed by a human review by the TASS organization.

Table 25 shows the polarity distribution of tweets in the collection, for both the training and test sets. As we can see, distributions are dissimilar between the two sets. This should arguably not be seen as a weakness of the corpus, but rather as a characteristic that is coherent with real-life settings, since the frequencies of the polarities of the tweets that are posted each day change depending on the topic. Regarding this issue, some studies (Brown and Levinson, 1987; Kennedy and Inkpen, 2006) highlight a general tendency of human language to positive classification, which could justify the presence of more positive reviews in training corpora.

The models proposed above are evaluated through two sets of experiments, in order to measure how the size of the training corpus can affect phenomena such as sparsity. In both cases, we perform a standard three class categorization considering positive (POS), negative (NEG) and without opinion (NONE) classes from the TASS corpus. In this case, neutral tweets will be discarded and strong positive and strong negative tweets will be included in the positive and negative classes, respectively. This means that performance will not be directly comparable to the systems which participated at the TASS workshop, where only classification into 4 and 6 categories was proposed. To overcome this limitation, additional experiments on 4 and 6 classes are included for the best performing models.

Category	#tweets training set	#tweets test set
POS+	1 652 (22.9%)	20 745 (34.1%)
POS	1 233 (17.1%)	1 488 (2.4%)
NEU	670 (9.3%)	1 305 (2.1%)
NEG	1 335 (18.5%)	11 287 (18.6%)
NEG+	847 (11.7%)	4 557 (7.5%)
NONE	1 482 (20.5%)	21 416 (35.2%)
Total	7 219(100.0%)	60 798 (100.0%)

Table 25: Frequency statistics of the TASS corpus

6.2.2 *Evaluation*

We try two different configurations to train and evaluate the proposed sets of features:

- *From small to large corpus*: This first set-up relies on the training set of the TASS corpus to build the models, and we evaluate them against

the test set. The training set of the TASS corpus only contains 6 549 tweets if we just consider those in the classes POS, NEG and NONE.

• *From large to small corpus*: We use the test set of the TASS to train the models, and we evaluate them against the training set. In order not to cause confusion, we refer to the test set as the reversed training set and the training set as the reversed test set. The aim of this experiment is to measure the effect of sparsity on the different models proposed. The size of the reversed training set is 59 493; considering positive, negative and none tweets, so it is around 10 times bigger than the original training set. We have also trained models using incremental parts of the reversed training set, to show how its size may affect to the accuracy of different perspectives. We are aware that the reversed training set can present some annotation errors, because it was made by pooling, followed by a human revision. We hypothesize that this will manifest itself in the form of a somewhat lower yield on the reversed test set, but not in the practical utility of the perspectives proposed. Optimization of models was made over this configuration, so we decided to split (fifty-fifty) the reversed test set into two parts: a development set, to analyze how properly combine different sets of features, and a test set to evaluate the real performance of selected models.

From small to large corpus configuration

Results are shown in Table 26. The model relying on lemmas obtains the best performance, followed by the pure bag-of-words model. Table 27 shows how the performance improves over the initial learning-based settings when features are used in conjunction. We obtained the best performance by creating a model which combines lemmas, psychometric properties, and the information provided by the unsupervised system. Specifically, the semantic orientation and the number of positive and negative words that appear in a tweet. We take the accuracy obtained by this combined model as a good indicator of what can be achieved without considering relations between words. We then test the effect of including syntactic information over this lexicon-based model, by adding generalized dependency triplets. We did not achieve any improvement incorporating syntactic features, following this experimental run, but Table 28 shows the results for some models which were able to improve performance when the collection is larger.

From large to small corpus configuration

Table 29 shows the results while Table 30 aims to show how their accuracy is improved when features are combined. As in the *from small to large* experiments, we obtain the best performing lexical model by creating a classifier which combines lemmas, psychometric proper-

Features	#features	pos-f1	neg-f1	none-f1	Accuracy
Lemmas (L)	755	**0.731**	**0.674**	**0.580**	**0.669**
Words (W)	851	0.701	0.655	0.557	0.645
Sentiment information (S)	3	0.641	0.576	0.575	0.600
Psychometric (P)	57	0.654	0.601	0.501	0.594
Fine-grained PoS-tags (FT)	86	0.611	0.561	0.474	0.559
Dependency types (D)	33	0.575	0.592	0.565	0.519
Bigrams of lemmas (BL)	998	0.592	0.565	0.295	0.514
Coarse-grained PoS-tags (CT)	16	0.552	0.489	0.440	0.504
Bigrams of words (BW)	915	0.573	0.528	0.204	0.480
Naive baseline	1	0.544	0.000	0.000	0.374

Table 26: Performance of the Spanish supervised model and the basic sets of features on the TASS corpus, following the *from small to large* setup: #features refers to the number of features of each model with an information gain greater than 0. POS-F1 , NEG-F1 and NONE-F1 refer to the value of F1 calculated for the positive, negative and none classes, respectively. Accuracy refers to the global accuracy, calculated over all the classes of tweets. *Naive baseline* is a trivial model, established by assigning all the instances to the majority class in the training set.

Features	#features	#pos-f1	neg-f1	none-f1	Accuracy
L ∪ P ∪ S	601	**0.765**	**0.702**	**0.609**	**0.700**
L ∪ P ∪ FT ∪ S	696	0.764	0.701	0.608	0.698
L ∪ P	598	0.749	0.688	0.592	0.684

Table 27: Performance of the Spanish supervised model on combining sets of lexical features on the TASS corpus, following the *from small to large* setup: lemmas (L) psychometric (P), fine-grained PoS-tags (FT), sentiment information (S). The symbol (∪) is used to represent concatenation of sets of features

Features	#features	#pos-f1	neg-f1	none-f1	Accuracy
L∪P∪S	601	**0.765**	**0.702**	**0.609**	**0.700**
L∪P∪S∪L$\xrightarrow{d}$P	1 102	0.756	0.695	0.600	0.692
L∪P∪S∪L$\xrightarrow{d}$CT	1 242	0.756	0.696	0.600	0.692
L∪P∪S∪L→P	1 131	0.757	0.697	0.600	0.692
L∪P∪S∪L→CT	1 244	0.712	0.696	0.600	0.691
L∪P∪S∪L→L	1 319	0.751	0.692	0.590	0.686

Table 28: Performance of the Spanish supervised model on incorporating generalized dependency features on the TASS corpus, following the *from small to large* setup. We use the notation $g(i,\zeta)del(\xrightarrow{d})g(j,\zeta)$ for representing sets of generalized dependency triplets. Also: lemmas (L) psychometric (P), coarse-grained PoS-tags (CT) , fine-grained PoS-tags (FT), dependency types (DT), sentiment information (S).

ties and the information provided by the unsupervised analyzer. This combined model is again taken as the base point from which to include syntactic information, in order to test the real effectiveness of generalized dependency features. The goal is to measure how relating terms, psychometric properties or part-of-speech information, by means of dependency parsing, can increase accuracy with respect to employing this knowledge in a purely lexical way. Table 31 illustrates some improvements obtained on accuracy when different generalized dependency triplets and the features of the best performing lexical model are used together. Given the number of possible combinations of generalized features, we only provide results for those that obtained some degree of improvement. We also include results for the best models that we achieved by combining several types of generalization. Asterisks indicate a statistically significant difference using chi-square tests. Unlike the configuration *from small to large*, in this case syntactic information is useful to improve performance. This suggests that although useful, generalized dependency triplets suffer from sparsity and a larger training set is needed to properly exploit this type of feature. We show below some cases where we believe that generalized dependency triplets were helpful to correct the polarity assigned by the best lexicon-based model on some difficult tweets:

- *'@Maropopins5:jajaja creo que es peor este que vi yo. Otro incunable ;)'* (*'@Paropopins5:hahaha I believe this one I saw it is worse. Another incunable ;)'*). The best-performing lexical model determined that this tweet is positive, while it was annotated as NEG in the TASS corpus. Although the model identifies the negative word *'worse'* it also recognizes the laugh *'hahaha'* and the emoticon *';)'* as positive terms and finally decides to take the tweet as POS. The main issue is that the lexical perspective does not differentiate between words forming part

of the core of the sentence and those simply offering auxiliary information. In this respect, the employment of generalized dependency triplets helps to take into account the syntactic structure of tweets in order to assign greater relevance to main syntactic functions such as the subject, direct object or subject complement, on which most of the meaning of the sentence relies. In this case, the term *'worse'* is the subject complement of the sentence, so the model considers the triplet (*'is'*, subject complement, *'worse'*). By backing off the head of this feature to their psychometric properties, the best-performing syntactic model matches it with generalized triplets such as (*Present time*, subject complement, *'worse'*) or (*Reference to other*, subject complement, *'worse'*), which are a priori negative, and finally classifies this tweet in its right category (NEG).

- *'Cansada de la familia Livela'* (*'I'm tired of Livela family'*). The lexical model classified this tweet as objective, due to the word *'tired'*, which when analyzed in isolation does not express any opinion, but simply describes a lack of energy. However, it is important to note the difference between saying *'tired'* and *'tired of'*. The syntactic model is able to correctly deal with the triplet (*'tired'*, prepositional object, *'of'*) and assign this feature to more general ones such as (*Sleep*, prepositional object, *'of'*) or (*Physical*, prepositional object, *'of'*) where Sleep and Physical are both psychometric properties. Using these generalized features is better than employing the original non generalized feature, because in addition to *'cansado de'* (*'tired of'*) they also encapsulate the meaning of similar Spanish phrases such as *'aburrido de'* (*'bored of'*) o *'harto de'* (*'sick of'*).

Features	#features	#pos-f1	neg-f1	none-f1	Accuracy
W	4 288	0.767	**0.691**	**0.622**	**0.702**
L	3 192	**0.769**	**0.691**	**0.622**	0.701
BL	9 066	0.731	0.657	0.575	0.659
BW	9 441	0.694	0.596	0.547	0.617
S	3	0.635	0.548	0.523	0.577
FT	148	0.603	0.548	0.513	0.560
P	63	0.595	0.576	0.513	0.559
D	40	0.553	0.455	0.502	0.511
CT	16	0.517	0.454	0.484	0.489
Naive baseline	1	0.611	0.000	0.000	0.440

Table 29: Performance of the Spanish supervised model and the basic sets of features on the TASS corpus, following the *from large to small* setup

Features	#features	#pos-f1	neg-f1	none-f1	Accuracy
L∪P∪FT∪S	3 031	**0.779**	**0.708**	**0.634**	**0.715**
L∪P∪S	2 881	**0.779**	0.701	**0.634**	0.713
L∪P	2 878	0.774	0.700	0.628	0.708

Table 30: Performance of the Spanish supervised model on combining sets of lexical features on the TASS corpus, following the *from large to small* setup. L∪P∪FT∪S obtain an small improvement over the L∪P∪S model at the test set, the L∪P∪S approach was taken as the starting point to incorporate syntactic information, since it obtained the best performance at the development set.

Features	#features	#pos-f1	neg-f1	none-f1	Accuracy
L∪P∪S	2 881	0.779	0.701	0.634	0.713
L∪P∪S∪L$\xrightarrow{d}$CT∪L→CT∪ L→FT∪P$\xrightarrow{d}$L∪CT→P*	25 996	**0.784**	**0.720**	0.635	**0.722**
L∪P∪S∪L→CT	7 660	0.782	0.713	**0.638**	0.718
L∪P∪S∪L$\xrightarrow{d}$CT	8 189	0.782	0.713	**0.638**	0.717
L∪P∪S∪L$\xrightarrow{d}$P	8 671	0.783	0.702	**0.638**	0.716
L∪P∪S∪L→L	11 057	0.779	0.706	0.635	0.714

Table 31: Performance of the Spanish supervised model on incorporating generalized dependency features on the TASS corpus, following the *from large to small* setup. The model marked with an '*' shows a statistically significant difference ($p < 0.05$) with respect to the L∪P∪S method.

Experiments on 4 and 6 classes

As we indicated in the experimental setup section, tweets in the TASS corpus are annotated with six labels and can thus be used to test performance on a more fine-grained scale of polarities. In this respect, Tables 32 and 33 present the performance for the most relevant feature models when they are used to classify polarity into 4 and 6 categories, respectively, using the *from small to large* configuration. Tables 34 and 35 show experimental results for 4 and 6 categories, respectively, this time according to the *from large to small* configuration.

Features	#features	#pos-f1	neu-f1	neg-f1	none-f1	Accuracy
L∪P∪S	428	**0.760**	**0.124**	**0.684**	**0.609**	**0.677**
L	485	0.715	0.086	0.641	0.568	0.636

Table 32: Performance of some supervised models on the TASS corpus, obtained from the *from small to large* setup, evaluated over 4 categories: positive, neutral, negative and none tweets.

Features	#features	pos+-f1	pos-f1	neu-f1	neg-f1	neg+-f1	none-f1	Accuracy
L∪P∪S	237	**0.697**	0.218	**0.158**	**0.534**	**0.535**	**0.646**	**0.586**
L	220	0.671	**0.239**	0.121	0.493	0.518	0.623	0.566

Table 33: Performance of some supervised models on the TASS corpus, obtained from the *from small to large* setup, evaluated over 6 categories: strong positive, positive, neutral, negative, strong negative and none tweets.

Features	#features	pos-f1	neu-f1	neg-f1	none-f1	Accuracy
$L \cup P \cup S \cup L \xrightarrow{d} CT \cup L \rightarrow CT \cup L \rightarrow FT \cup P \xrightarrow{d} L \cup CT \rightarrow P^*$	17 876	**0.748**	**0.178**	0.647	**0.603**	**0.643**
L∪P∪S	2 127	0.739	0.098	**0.652**	0.592	0.639
L	2 366	0.728	0.118	0.650	0.587	0.633

Table 34: Performance of some supervised models on the TASS corpus, obtained from the *from large to small* setup, evaluated over 4 categories: positive, neutral, negative and none tweets. The model marked with an '*' shows a statistically significant difference ($p<0.01$) with respect to the L∪P∪S method.

6.2.3 *Discussion of the features*

From Tables 26 and 29, which show the performance of the basic feature models with both the *from small to large* and *from large to small*

Features	#features	pos+-f1	pos-f1	neu-f1	neg-f1	neg+-f1	none-f1	Accuracy
$L \cup P \cup S \cup L \xrightarrow{d} CT \cup L \rightarrow CT \cup L \rightarrow FT \cup P \xrightarrow{d} L \cup CT \rightarrow P^*$	12 671	**0.669**	**0.225**	**0.154**	0.484	**0.495**	**0.598**	**0.525**
$L \cup P \cup S$	1 649	0.652	0.141	0.093	**0.485**	0.465	0.578	0.507
L	1 726	0.649	0.157	0.093	0.469	0.479	0.578	0.504

Table 35: Performance of some relevant models obtained from the *from large to small* setup, evaluated over 6 categories: strong positive, positive, neutral, negative, strong negative and none tweets. The model marked with an '*' shows a statistically significant difference ($p<0.01$) with respect to the $L \cup P \cup S$ method.

configurations, the tendency with respect to accuracy remains very similar in both runs. The model simply fed with lemmas seems to be the most successful set of features, followed by the approach using words as features. In particular, in the *from small to large* run, the use of lemmas clearly outperforms the use of words. This shows the need to apply some type of normalization of Spanish words, reducing the rich morphology of this language, but keeping the meaning of words. With respect to the *from large to small* run, both lemmas and words obtain virtually the same accuracy, although lemmas employ a much lower number of features. It is important to note that a model based on only words instead of on lemmas, implicitly captures features such as gender or number, which are good features by themselves, as we will discuss below. Thus, a model based on words contains, to a certain extent, analogous information to that included in a combined model of bag-of-lemmas and fine part-of-speech tags. Models based on bi-grams show a low performance, probably due to the sparsity of these features in a small training set. The psychometric approach also achieves a decent performance, strengthening the importance of taking semantic approaches as a starting point. Table 36 illustrates the top 5 features for these three approaches, based on their information gain, taking the *from large to small* configuration. The pair '*the/he*' would correspond to '*el/él*' in Spanish language. Actually, the second best discriminative was just '*el*'. However, as we commented previously, Spanish users often ignore acute accents when writing in web environments and furthermore articles are often omitted in microtexts. Therefore, we hypothesize the form '*el*' many times really refers to '*él*'. In this respect, third person pronouns are often good indicators of objective texts, since informative texts often present a distance from the sender, whilst opinions are more frequently expressed with first or second person pronouns.

An interesting finding is the accuracy obtained by only using part-of-speech tags. Although it hardly provides any explicit semantic information, the fine-grained part-of-speech tags model obtains an accuracy similar to the psychometric approach. This suggests that

Ranking	Lemmas	Words	Psychometric
1	Positive emoticon	Positive emoticon	Positive-emoticon
2	the/he	!	Affective
3	!	Not	Negative-emotion
4	Thanks	That	Positive-emotion
5	Not	Thanks	Article

Table 36: Top 5 discriminative features for the basic sets of features from the TASS training corpus, according to information gain and following the *from small to large* setup

features such as gender, number or some word categories (e.g., conjunctions) can be good classifiers in themselves. Table 37 shows the ranking of the top fine-grained PoS tags, according to their information gain in the training set, which reinforces this hypothesis. Labels such as the close exclamation mark, or the artificial emoticon-tag, are two of the most discriminative features, probably because they are good indicators of subjective tweets. In Spanish there exists also an open exclamation mark '¡', conventionally used to mark the beginning of an exclamation, but users often ignore it in web environments. The occurrence of the tag subordinating conjunction in the top five of the best part-of-speech features suggests the importance of identifying adversative subordinate clauses, as we have pointed out previously. Subordinating constructions often compare and oppose arguments, which represents a good point to identify subjective texts. The fine-grained PoS-tag *Verb 3rd person singular present indicative* is intuitively a good indicator of objective texts, as has been noted by other authors (Pak and Paroubek, 2010; Spencer and Uchyigit, 2012): people giving an opinion tend to use first person pronouns, because they are probably talking about something that happened to them; but the same is usually not true for people who are merely reporting on a fact, where third person pronouns are more frequent. In Spanish, subject pronouns are usually eliminated, since inflected verb forms provide us with the information needed to determine the number of the subject, which can be helpful to differentiate between subjective and objective texts, as we have just described.

Dependency types, which represent the syntactic functions present in a tweet, seem not to be very helpful in themselves.

Tables 27 and 30 show how we can improve performance in an effective way by combining different sets of basic features, obtaining a better lexicon-based model. Combined models which incorporate unsupervised sentiment information, are not purely lexicon-based, since our semantic orientation analyzer uses heuristic syntactic rules. For both runs, the classifier whose features are the lemmas, psychometric properties, semantic orientation and the number of positive and nega-

Ranking	Feature
1	Close exclamation mark
2	Verb 3 person singular present indicative
3	Negative adverb
4	Emoticon-tag (artificial tag)
5	Subordinating conjunction

Table 37: Top 5 discriminative fine-grained part-of-speech tags, according to information gain following the *from small to large* setup

tive words that appear in a tweet achieved the best performance. This allows us to establish a ceiling of effectiveness for dealing with terms in an isolated way. Moreover, with this model we reduced the number of features with an information gain greater than zero with respect to the best basic approach, the bag-of-lemmas perspective. Other lexicon-based models which add linguistic information such as part-of-speech tags or dependency types did not increase the accuracy (difference not statistically significant. $p < 0.10$). Table 38 shows some of the most discriminative features for the best combined model which does not take into account generalized dependency triplets. The information provided by SSSA seems to be highly relevant, validating the utility of that approach. The most discriminative lemma appears in the eighth position, although lemmas were the best approach when they were considered in isolation.

Ranking	Feature	Provided by
1	Semantic orientation	The unsupervised system
2	Positive emotion	Psychometric approach
3	#positive words	The unsupervised system
4	#negative words	The unsupervised system
5	Affective	Psychometric approach
8	Positive emoticon	Lemmas approach

Table 38: Relevant discriminative features when combining lemmas, psychometric properties and the information provided by our unsupervised system, according to information gain and following the *from large to small* setup

Tables 28 and 31 reflect the effect on performance when syntactic information are provided in the form of generalized dependency triplets; both for the *from small to large* and *from large to small* configurations. With respect to the *from small to large* runs, generalized dependency triplets do not improve the performance over the best

lexical model. This is due to the high sparsity of this type of feature and the relatively small size of the training corpus, which is not even able to successfully exploit a model based on a bag-of-lemmas, as we have seen previously. On the other hand, in the *from large to small* experiments, syntactic information are helpful to improve performance over purely lexical models. If we incorporate different types of generalized dependency triplets over the lexical model we obtain small improvements, but when several of these features are jointly aggregated we obtain an even higher accuracy. It is important to note that the best models were mainly obtained by including features which carry out a high level of generalization on the dependent node, contradicting the approach proposed by Joshi and Penstein-Rosé (2009), who suggested that it is better to generalize the head node. However, when generalized dependency triplets were evaluated in isolation, performing a higher generalization on the head node was more appropriate.

Table 39 presents a sample of interesting features for the model which obtained the best performance on the *from large to small* configuration. Some of these features show how Spanish users relate terms according to the frame of mind of society at large. For example, the term *police* appears directly associated with the psychometric category *negative emotion*, probably due to the strikes and demonstrations occurring in Spain during the period in which the tweets were collected. Along the same lines, Spanish users relate the word *economy* with *negative emotion*. Picking topic terms is pointed out as a risk on building supervised sentiment classifiers. Given a training corpus, if a topic word such as *'economy'* or *'police'* appears mostly in one class, those words should not be considered for analyzing new tweets, due to the bias of the training set. Our approach is no exception to this limitation, because we are including unigrams of lemmas. However, the use of composite generalized features can diminish this phenomena, since we are able to relate topic words with psychological properties, which are fine complements for topic words, as is shown at the examples of the Table 39. Moreover, we realized that generalized dependency triplets were able to catch, to a certain extent, the discourse structure on Twitter. As we can see in the same table, to classify the polarity of a tweet the use of the word *'thanks'* at the end of the sentence seems to be more relevant than explicitly thanking somebody (shown by the feature $\mathtt{thank} \rightarrow \mathtt{proper\ name}$).

Models with a small number of features, such as psychometric or fine-grained part-of-speech tags, do not benefit from a larger training set, as expected. The same is not true for more complex models, which clearly improve their performance with a larger training corpus. Figure 9 illustrates how the size of the training set, following the *from large to small*, setup affects the performance of some of the models showed above. The X axis indicates the percentage of the reversed

Ranking	Feature	Provided by
48	Thanks → punctuation mark	Lemmas → Coarse tag
349	Thanks → proper name	Lemmas → Coarse tag
447	noun → Anxiety	Coarse tag → Pyschometric
6863	Negative emotion $\xrightarrow{s.a}$ police	Psychometric $\xrightarrow{dp}$ Lemmas
19417	Negative emotion $\xrightarrow{suj}$ economy	Psychometric $\xrightarrow{dp}$ Lemmas
19421	Reference to other $\xrightarrow{suj}$ Austerity	Psychometric $\xrightarrow{dp}$ Lemmas

Table 39: Relevant generalized dependency features for the best performing model, according to information gain and following the *from large to small* setup

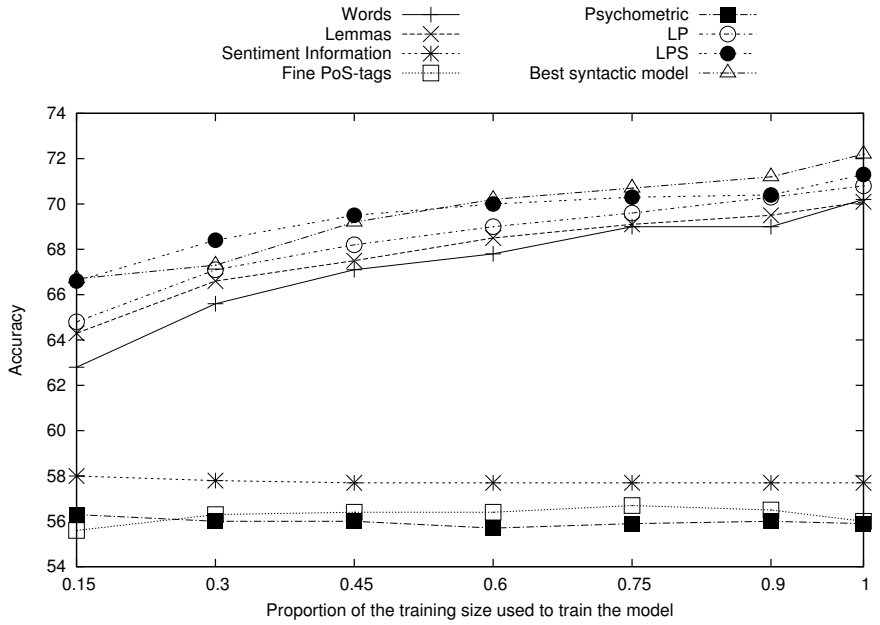


Figure 9: Performance following the *from large to small* setup for different models, using incremental pieces of the training collection to build them

training set employed to build the model, and the Y axis corresponds the accuracy. It is important to remark that the lemmas model outperforms the words model when the training collection is small, but the performance between the two approaches converges as the training set grows. When the training set is not large enough (45% of the corpus equals 26 770 tweets) generalized triplets are not helpful to improve the accuracy of the model composed by lemmas, psychometric and sentiment information, which is the best one in these cases. But for larger training sets, the generalized syntactic model outperforms the rest of perspectives.

Regarding to the results over 4 and 6 classes, the tendency of the performance seems to be coherent with respect to the experiments over 3 classes; syntactic approaches once again outperform lexical

models, and thus the discussion provided above could also be extrapolated to these runs. In all experiments the best syntactic model obtained a statistically significant difference with respect to the best lexical model, when the training set is large enough, validating the utility of generalized dependency triplets over a wide range of polarity categories.

6.3 CONCLUSION

This chapter focused on predicting the sentiment of tweets written in the Spanish language, by means of supervised linguistics-based methods. We provided an evaluation which ranged from standard learning-based methods to shallow and deep linguistic approaches. The main contribution of the chapter relies on testing how relating lexical, syntactic, psychological and semantic information affects polarity classification on tweets. To the best of our knowledge, this was the first work which performed a wide evaluation of the effectiveness of using these features, both in isolation and in combination, on a corpus of Twitter messages.

With respect to syntactic features, we rely on a more relaxed variant of the generalized dependency triplets proposed by Joshi and Penstein-Rosé (2009) to identify opinionated sentences. We adapt the method to perform polarity classification on tweets, enriching their angle by considering various levels of generalization, ranging from part-of-speech to psychological and semantic abstraction. The utility of syntax on sentiment analysis is a widely discussed issue, but it has often been focused on long and specific domain reviews. To the best of our knowledge, this is also the first work which studies the effect of dependency parsing on Spanish tweets. Empirical results suggest that non-syntactic approaches obtain a better performance when the training set is small, but as the size of the training corpus grows, the incorporation of generalized dependency triplets helps to improve accuracy over the purely lexical perspectives.

7

A PROPOSAL TO MUTILINGUAL SUPERVISED SENTIMENT ANALYSIS

In the previous chapter, we built a supervised machine learning model that combined linguistic (lexical, syntactic and semantic) information to determine the polarity of Spanish tweets. In the line of Chapter 5, we now explore how it is possible to add multilingual functionalities to our monolingual supervised model, initially created for Spanish.

In particular, we compare three techniques: (1) a multilingual model trained on a multilingual dataset, obtained by fusing existing monolingual resources, that does not need any language recognition step, (2) a dual monolingual model with perfect language detection on monolingual texts and (3) a monolingual model that acts based on the decision provided by a language identification tool. The aim of this chapter is not to introduce a new sentiment analysis architecture, but to show how current state-of-the-art supervised approaches can successfully address situations where monolingual, multilingual and code-switching texts occur.

The techniques are evaluated on monolingual, synthetic multilingual and code-switching[1] corpora of English and Spanish tweets. In the latter case, we introduce the first code-switching Twitter corpus with sentiment labels. The samples are labeled according to two well-known scales, already used in this book: the *sentistrength* scale and the trinary scale.

The experimental results show the robustness of the multilingual approach (1) and also that it outperforms the monolingual models on some monolingual datasets.

7.1 DESCRIPTION

The approach followed in this chapter is similar to the one introduced in Chapter 6. We are relying on similar sets of features (see §6.1.4) and also on standard machine learning algorithms. In particular, in this chapter we relied on an L2-regularized logistic regression (Fan et al., 2008). In general, linear classifiers have provided state-of-the-art performance since early research on SA (Mohammad, Kiritchenko, and Zhu, 2013; Paltoglou and Thelwall, 2010) and in particular, logistic regression is a good fit for this task (Jurafsky and J.H, 2016).

The novelty of the chapter comes from adding multilingual functionalities to the model, so it can analyze different languages. In particular we focus on a bilingual environment containing English

1 Code-switching texts are those that contain terms in two or more different languages.

and Spanish texts. Following such sets of features, we propose three different approaches:

1. *Multilingual approach (en-es model)*: we have only one model that works on both Spanish and English texts. The *en* and *es* training and development corpora are merged to train a unique *en-es* sentiment classifier.

2. *Dual monolingual approach (en and es models)*: We have two monolingual models, one for Spanish and another for English. This approach represents the ideal (unrealistic) case where the language of the text is known in advance and the right model is executed. Each language model is trained and tuned on a monolingual corpus.

3. *Monolingual pipeline with language detection (pipeline approach)*: We also have two monolingual models, one for Spanish and the other for English, but in this approach we first identify the language of a text through the `langid.py` (Lui and Baldwin, 2012) language detection software, where the output language set was constrained to Spanish and English to make sure every tweet is classified and guarantee a fair comparison with the other approaches. The training was done in the same way as in the monolingual approach, as we know the language of the texts. `langid.py` is only needed for evaluation, not for training. Experiments are performed considering the following pipeline: The language is predicted; then, the corresponding monolingual classifier is called; and finally the outputs are joined to compare them to the gold standard.

7.2 CODE-SWITCHING

Code-switching texts are those that contain terms in two or more different languages, and they occur increasingly often in social media. The aim of this section is to provide a corpus (EN-ES-CS) to the research community to evaluate the performance of sentiment classification techniques on this complex multilingual environment, proposing an English-Spanish corpus of tweets with code-switching.

To build the EN-ES-CS corpus, we take as starting point the collection presented in Solorio et al. (2014), a workshop on language detection on code-switching tweets, where the goal was to apply language identification at the word level. The organizers proposed four code-switching language detection challenges from different language families: Spanish-English, Nepali-English, Mandarin-English and Modern Standard Arabic-Arabic dialects. They made the training corpora available to the research community, together with a small tuning collection, but no test set was released at the moment this research was carried out.

For building our resource, we just considered the Spanish-English training set (originally 11 400 tweets). As a first step, we removed all

the non code-switching texts, i.e. those where all the words belonged to the same language, obtaining a filtered collection of 3 062 tweets. A number of different types of tweets can be found in the corpus, as shown in Example 21.

EXAMPLE 21 (Code-switching samples from the corpus). Different code-switching samples from the corpus presented in this chapter that illustrate different options to express sentiment in this kind of texts. The double underline represents the English text, the simple underline the Spanish phrases, and no underline illustrates language independent symbols.

- Tweets that show (even opposite) sentiment in both languages, e.g. *'Tan bien que ivan las cosas... im so lost what did i do?!'*.
- Tweets where the sentiment is just in the English side of the tweet, e.g. *'I legitally screamed!!!! No fue una si no dos!!!'*.
- Tweets where the sentiment is just in the Spanish side of the tweet, e.g. *'This house da miedo'*.
- Tweets where the sentiment relies on language-independent symbols, e.g. *'Wow no lo puedo creer? -.-'*.

□

Tweets were sent to three annotators fluent both in Spanish and English, who were asked to annotate them according to the *sentistrength* criteria (Thelwall et al., 2010) and the Wiebe, Wilson, and Cardie (2005) annotation style. *SentiStrength* is a dual-score sentiment labeling strategy where each text is given two scores between 1 and 5: one indicating the positive strength (*ps*) of the tweet and the second one indicating its negative strength (*ns*). See Chapter 3 for a detailed description. For example, *'I love you, but I hate you'* would have both a strong positive and negative sentiment. For inter-annotator agreement we relied on Krippendorff's alpha coefficient (Hayes and Krippendorff, 2007), obtaining an agreement from 0.629 to 0.664 for negative sentiment and 0.500 to 0.693 for positive sentiment.

Table 40 shows the frequency distribution of the *sentistrength* scores and how annotators tend to often find slight levels of subjectivity, while highly subjective tweets tend to be less frequent.[2]

The results are coherent with other corpora annotated according to these criteria such as (Thelwall et al., 2010) or the Spanish *sentistrength* corpus presented in Chapter 3. The corpus was observed to be especially noisy, with many grammatical errors occurring in each tweet. Additionally, a predominant use of English was detected. We believe this is because the Solorio et al. (2014) corpus was collected

2 Words such as *'good'* or *'bad'* tend to be more often used than *'spectacular'* or *'horrible'*, that are reserved for more special occasions.

Positive	%tweets	Negative	%tweets
1	63,26	1	69,42
2	26,58	2	19,59
3	7,54	3	8,43
4	2,35	4	2,15
5	0,26	5	0,04

Table 40: Frequency distribution of the *sentistrength* scores on the EN-ES-CS CORPUS

by downloading tweets posted by people from Texas and California, where English is the primary language. Table 41 reflects these particularities.[3] In total, our collection contains 24 758 English terms, with 5 565 unique words, where 3 576 of them turned out to be out-of-vocabulary (OOV) words. Spanish is the minority language in the corpus, with 16 174 occurrences of terms and only 5 033 unique words, although with a larger percentage of OOV words. We also ran a language detection system, `langid.py`, resulting in 59.29% of tweets being predicted as English tweets.

Finally, there is also a nearly ubiquitous use of subjective clauses and abbreviations, especially *'lol'* and *'lmao'*, whose sentiment was considered a controversial issue by the annotators. It is interesting to point out that the presence of these clues was also used sometimes as a part of a negative message (i. e. *'He is so stupid, lmao'*), without any positive connotation. We believe this could have been one of the reasons why the inter-annotator agreement was lower for positive than for negative scores.

Language	Word occurrences	Unique words	oov words
English	24 758	5 565	3 576
Spanish	16 174	5 033	3 714

Table 41: Word statistics by language on the EN-ES-CS corpus. Symbols like numbers or punctuation marks were considered language independent by Solorio et al. (2014)

Table 42 shows some of the most common terms observed in our corpus that usually have sentiment associated, confirming the tendency of the users to employ subjective interjections coming from

3 The words present in McDonald et al. (2013)'s English and Spanish treebanks were taken as our dictionaries. To know the language of each word of the corpus, we rely on Solorio et al. (2014)'s annotations.

English. It is also important to note that the Spanish terms usually involve Mexican Spanish varieties, so specific resources from these might be needed to improve performance on the Spanish phrase sentiment classification.

English term	Occurrences	Spanish term	Occurrences
'lol'	474	*'bien'*	61
'like'	170	*'jajaja'*	29
'lmao'	122	*'mejor'*	28
'haha'	67	*'pinche'*	25
'good'	64	*'quiero'*	22
'love'	47	*'kiero'*	19
'shit'	47	*'jaja'*	18
'fuck'	42	*'guey'*	15
'better'	29	*'pedo'*	14

Table 42: Occurrences of some of the most common subjective terms for English and Spanish in the EN-ES-CS corpus

Additional labeling

A second labeling strategy is also provided for the code-switching corpus. After averaging the annotator scores, we applied a transformation to the *de facto* standard polarity classes (positive, neutral and negative) (Nakov et al., 2013; Nakov et al., 2016a; Rosenthal et al., 2014). If $ps > ns$ then the tweet was considered *positive*. If $ps < ns$ then the tweet was considered *negative*. Otherwise, it was taken as *neutral*.[4] After the conversion, we obtained a collection where the *positive* class represents 31.45% of the corpus, the *negative* one represents 25.67% and with a 42.88% of neutral tweets. This frequency distribution is also close to that of other widely used Twitter corpora (Rosenthal et al., 2014). Both versions of the EN-ES-CS CORPUS can be obtained at `http://www.grupolys.org/software/CS-CORPORA/`. We have tagged the corpus following different strategies in order to provide a richer resource, giving users the opportunity to select the tagging scheme that best suits their needs.

The format of the corpus labeled according to *sentistrength* is:

`ps \t ns \t tweetid \t text`

and the format of the corpus labeled according to the trinary scale is:

`polarity \t tweetid \t text`.

4 Neutral tweets can be either totally objective or mixing positive and negative sentiment with the same strength. However, the latter case turned out to be very uncommon.

where for each tweet, `ps` refers to its positive strength, `ns` to its negative strength, `tweetid` to its unique identifier, `text` to its contents, `polarity` to its polarity class and `\t` is used to represent a tab character.

7.3 EXPERIMENTS

7.3.1 *Datasets*

The corpora used to evaluate the proposed approaches are:

1. *SemEval 2014 task B corpus* (Rosenthal et al., 2014): A set of English tweets[5] split into training (8 200 tweets), development (1 416) and test sets[6] (5 752). Each tweet was manually classified as positive (POS), objective (NONE) or negative (NEG).

2. TASS *corpus* (Román et al., 2015): A corpus of Spanish tweets containing a training set of 7 219 tweets. We split it into a new training and a development set (80:20). Two different test sets are provided: (1) a *general test set* of 60 798 tweets that was made by pooling and (2) a small test set of 1 000 manually labeled tweets, named *1K test set*. The tweets are labeled with positive (POS), objective (NONE), negative (NEG) and mixed (NEU), but in this study the NEU class was treated as NONE, following the same criteria as in SemEval 2014.

3. *Multilingual corpora resulting from merging SemEval 2014 and* TASS *corpora*. These two test sets were merged to create two synthetic multilingual corpora: (1) SemEval 2014 + TASS 1K (English is the majority language) and (2) SemEval 2014 + TASS general (Spanish is the majority language). The unbalanced sizes of the test sets result in a higher performance when correctly classifying the majority language. We do not consider this as a methodological problem, but rather as a challenge of monitoring social networks in real environments, where the number of tweets in each language is not necessarily balanced.

4. *The code-switching corpus described in §7.2.*

7.3.2 *Evaluation*

We show below the performance of each model in each of the four proposed configurations: (1) an English monolingual corpus, (2) a

5 Due to Twitter restrictions some of the tweets are no longer available, so the corpus statistics may vary slightly from those of other researchers that used the corpus.

6 It also contained short texts coming from SMS and messages from LiveJournal, which we removed as they are outside the scope of this study.

Spanish monolingual corpus, (3) a multilingual corpus which combines the two monolingual collections and (4) the code-switching (Spanish-English) corpus presented in §7.2.

Table 43 shows the performance of the three models on the SemEval English monolingual test set. The differences between the monolingual model and the monolingual pipeline with language detection are tiny. This is due to the high performance of `langid.py` on this corpus, where only 6 tweets were misclassified as Spanish tweets. In spite of this issue, the *en-es* classifier performs very competitively on the English monolingual test sets, with differences with respect to the *en* model ranging from 0.2 to 1.05 percentage points in terms of accuracy. With certain sets of features, the multilingual model even outperforms both monolingual models, reinforcing the validity of this approach.

Features	f1			Accuracy		
	en	pipe	en-es	en	pipe	en-es
Words (W)	**65.8**	65.7	65.4	**66.7**	**66.7**	66.2
Lemmas (L)	**65.8**	**65.8**	65.7	**66.7**	**66.7**	66.5
Psychometric (P)	**61.3**	**61.3**	60.2	**62.5**	**62.5**	61.5
PoS-tags (T)	48.0	48.0	**49.5**	51.8	51.8	**52.0**
Bigrams of W (BW)	59.1	59.1	**60.2**	61.0	61.0	**61.5**
Bigrams of L (BL)	**59.9**	**59.9**	**59.9**	**61.8**	**61.8**	61.3
Bigrams of P (BP)	**60.6**	**60.6**	59.8	**61.3**	**61.3**	60.4
W→W	53.1	53.1	**55.8**	56.4	56.4	**57.8**
L→L	56.0	56.0	**57.2**	58.7	58.7	**59.2**
P→P	**57.4**	**57.4**	56.9	**58.3**	58.2	57.6
W∪P∪T	68.0	**69.0**	68.2	68.5	**68.6**	**68.6**
L∪P∪T	**68.0**	67.8	67.9	**68.4**	**68.4**	68.3
W∪P	68.2	**68.3**	68.1	**68.7**	**68.7**	68.5
L∪P	**68.0**	**68.0**	67.8	**68.6**	68.5	68.3

Table 43: Performance on the SemEval 2014 test set by the monolingual, language-detection and multilingual models. We evaluate the English monolingual approach (*en*), the monolingual pipeline with language detection (*pipe*) and the multilingual approach (*en-es*). For each row, the best values of F1 and accuracy are shown in boldface.

With respect to the evaluation on the Spanish monolingual corpora, results on the TASS corpora are shown in Table 44, including results on both the general and the TASS -1K test sets. With respect to the evaluation on the TASS and TASS -1k corpora the *es* model obtains the best

results, followed by the *pipe* and the *en-es* models. In the TASS -1k test set, the language detection system misclassified 17 of the manually labeled tweets, and the impact of the monolingual model with language detection is also small. Results obtained on the TASS general set give us more information, since a significant number of tweets from this collection (842) were classified as English tweets. Some of these tweets actually were short phrases in English, some presented code-switching and some others were simply misclassified. Under this configuration, the multilingual model outperforms monolingual models with most of the proposed features. This suggests that multilingual models present advantages when messages in different languages need to be analyzed.

Features	1K test set						General test set					
	f1			Accuracy			f1			Accuracy		
	es	pipe	en-es	es	pipe	en-es	es	pipe	en-es	es	pipe	en-es
W	**58.2**	**58.2**	54.6	**56.6**	56.5	54.6	64.1	64.1	**64.4**	64.4	64.4	**64.6**
L	57.9	57.8	**58.2**	56.4	56.3	**56.6**	64.2	64.2	**64.3**	64.5	64.5	**64.6**
P	**56.1**	**56.1**	53.1	**54.7**	**54.7**	53.1	58.5	58.4	**59.3**	58.8	58.7	**59.5**
T	**49.4**	49.3	41.2	**48.9**	48.8	41.7	**49.3**	**49.3**	45.9	49.4	**49.5**	47.7
BW	**54.4**	54.2	53.9	**52.9**	52.7	52.1	58.2	58.3	**58.9**	58.4	58.4	**58.7**
BL	**55.5**	**55.4**	54.3	**54.0**	53.9	52.2	58.6	58.6	**59.3**	58.7	58.7	**59.3**
BP	47.6	47.6	**48.7**	46.0	46.0	**47.0**	51.3	51.2	**53.2**	51.3	51.3	**53.2**
W→W	**53.7**	53.5	46.7	**52.4**	52.2	44.6	54.0	54.2	**54.8**	54.2	54.4	**55.0**
L→L	**55.8**	**55.8**	48.4	**54.4**	**54.4**	46.3	55.9	55.9	**56.4**	56.1	56.1	**56.4**
P→P	**47.5**	**47.5**	**47.5**	45.8	**45.8**	47.5	50.0	50.0	**52.3**	50.0	49.4	**52.3**
W∪P∪T	61.5	**61.6**	60.8	**60.0**	59.9	59.1	**66.1**	66.0	**66.1**	**66.4**	66.3	66.3
L∪P∪T	**62.7**	**62.7**	60.8	**61.4**	**61.4**	59.2	65.8	65.7	**65.9**	**66.2**	66.1	66.1
W∪P	60.8	60.8	**61.2**	59.1	59.2	**59.6**	65.9	65.9	**66.0**	66.3	66.2	**66.3**
L∪P	61.3	**61.4**	60.9	59.8	**59.9**	59.3	65.6	65.6	**65.7**	**66.0**	65.9	65.9

Table 44: Performance on the TASS test sets by the monolingual, language-detection and multilingual models. We evaluate the Spanish monolingual approach (*es*), the monolingual pipeline with language detection (*pipe*) and the multilingual approach (*en-es*). For each row, the best values of F1 and accuracy are shown in boldface.

Table 45 shows the performance both of the multilingual approach and the monolingual pipeline with language detection when analyzing texts in different languages. On the one hand, the results show that using a multilingual model is the best option when Spanish is the majority language, probably due to a high presence of English words in Spanish tweets. On the other hand, combining monolingual models with language detection is the best-performing approach when English is the majority language. The English corpus contains only a few Spanish terms, suggesting that the advantages of having a multilingual model cannot be exploited under this configuration.

Features	SemEval+tass -1K				SemEval+tass -general			
	f1		Accuracy		f1		Accuracy	
	pipe	en-es	pipe	en-es	pipe	en-es	pipe	en-es
W	**64.5**	63.7	**64.9**	64.2	64.3	**64.5**	64.6	**64.7**
L	**64.5**	**64.5**	**65.0**	64.8	64.3	**64.4**	**64.7**	**64.7**
P	**60.5**	59.1	**61.2**	60.0	58.7	**59.4**	59.0	**59.7**
T	48.1	**49.2**	**51.3**	50.2	**49.2**	46.2	**49.7**	48.1
BW	58.3	**59.2**	59.6	**59.8**	58.3	**59.0**	58.6	**58.9**
BL	**59.2**	59.0	**60.4**	59.7	58.7	**59.4**	59.0	**59.5**
BP	58.6	**58.8**	**58.7**	58.1	52.0	**53.8**	52.2	**53.9**
W→W	53.1	**54.4**	**55.7**	55.5	54.1	**54.9**	54.6	**55.2**
L→L	**55.9**	55.8	**57.9**	56.9	55.9	**56.5**	56.3	**56.6**
P→P	**55.8**	55.5	**56.1**	55.8	50.6	**52.7**	50.3	**52.8**
W∪P∪T	**67.8**	67.0	**67.1**	66.9	66.2	**66.3**	**66.5**	**66.5**
L∪P∪T	**67.0**	66.8	**67.2**	66.8	65.9	**66.1**	**66.3**	**66.3**
W∪P	**67.1**	67.0	**67.1**	67.0	66.1	**66.2**	66.4	**66.5**
L∪P	**66.9**	66.7	**67.0**	66.8	65.8	65.9	**66.1**	**66.1**

Table 45: Performance on the multilingual test set by the monolingual, language-detection and multilingual models. The first group of two columns represents the performance of the synthetic dataset SemEval+TASS -1k (English is the majority language) and the second group of two columns represents the performance on the dataset SemEval+TASS general (Spanish is the majority language). For each row, the best values of F1 and accuracy are shown in boldface.

Finally, Table 46 shows the performance of the three proposed approaches on the code-switching test set. The accuracy obtained by the proposed models on this corpus is lower than on the monolingual corpora. This suggests that analyzing subjectivity on tweets with code switching presents additional challenges. The best accuracy (59.34%) is obtained by the *en-es* model using lemmas and psychometric properties as features. In general terms, atomic sets of features such as words, psychometric properties or lemmatization, and their combinations, perform competitively under the *en-es* configuration. The tendency remains when the atomic sets of features are combined, outperforming the monolingual approaches in most cases.

The pipeline model performs worse on the code-switching test set than the multilingual one for most of the sets of features. These results, together with those obtained on the monolingual corpora, indicate that a multilingual approach like the one proposed in this chapter is more robust on environments containing code-switching tweets and tweets in different languages. The *es* model performs poorly,

	f1				Accuracy			
Features	**en**	**es**	**pipe**	**en-es**	**en**	**es**	**pipe**	**en-es**
W	**54.2**	45.2	51.6	54.1	**55.7**	47.7	52.7	54.89
L	54.3	46.2	51.9	**55.7**	55.9	48.9	53.0	**56.4**
P	52.2	40.8	50.0	**53.3**	53.0	43.6	50.7	**53.7**
T	38.5	34.4	40.2	**39.6**	**45.1**	39.3	44.7	43.2
BW	49.3	45.1	48.5	**51.9**	54.3	47.5	51.7	**54.3**
BL	50.1	46.4	49.1	**51.4**	**55.0**	48.9	52.2	53.6
BP	**47.7**	37.3	45.2	46.8	**49.5**	40.5	46.1	46.9
W→W	46.6	30.2	43.1	**47.1**	**52.6**	36.5	46.0	50.7
L→L	47.4	42.4	45.6	**47.8**	**53.0**	44.7	49.0	50.4
P→P	**46.2**	36.2	44.5	45.6	**48.1**	40.6	45.7	46.0
W∪P∪T	58.3	47.1	56.1	**58.5**	**59.2**	48.3	56.5	58.5
L∪P∪T	57.7	48.9	55.6	**58.6**	58.6	49.7	56.1	**59.1**
W∪P	58.0	48.4	55.9	**58.8**	58.7	49.9	56.4	**58.8**
L∪P	58.2	49.3	55.6	**58.9**	58.9	50.8	56.1	**59.3**

Table 46: Performance on the code-switching set by the monolingual, language-detection and multilingual models. For each row, the best values of F1 and accuracy are shown in boldface.

probably due to the smaller presence of Spanish words in the corpus. The annotators also noticed that Spanish terms present a larger frequency of grammatical errors than the English ones. Surprisingly, the *en* model performed really well in many of the cases. We hypothesize this is due to the higher presence of English phrases, which made it possible to extract the sentiment of the texts in many cases.

Experimental results allow us to conclude that the multilingual models proposed in this work are a competitive option when applying polarity classification to a medium where messages in different languages might occur. The results are coherent across different languages and corpora, and also robust on a number of sets of features. In this respect, for contextual features the performance was low in all cases, due to the small size of the training corpus employed. In Chapter 6 we explained how features of this kind become useful when the training data becomes larger.

7.4 CONCLUSION

In this chapter, we have compared different machine learning approaches to perform multilingual polarity classification in three different environments: (1) where monolingual tweets are evaluated

separately, (2) where texts in different languages need to be analyzed and (3) where code-switching texts occurred. To evaluate scenario (3), we have presented together with this chapter the first code-switching Twitter corpus for multilingual sentiment analysis, composed of tweets that merge English and Spanish terms.

The proposed approaches were: (a) a multilingual model trained on a corpus that fuses two monolingual corpora, according to level 2 (Situation Refinement) of Information Fusion techniques to the Sentiment Analysis pipeline, described by Balazs and Velásquez (2016), (b) a dual monolingual model and (c) a simple pipeline which used language identification techniques to determine the language of unseen texts.

Experimental results reinforce the robustness of the multilingual approach under the three configurations. The results obtained by this model on the monolingual corpora are similar to those obtained by the corresponding monolingual approaches (i.e. we can teach a supervised model an additional language without significant loss of performance). The results also show that neither monolingual nor multilingual approaches based on language detection are optimal to deal with code-switching texts, posing new challenges to sentiment analysis on this kind of texts.

Part IV

APPLICATIONS

8

TOPIC CLASSIFICATION

In Parts ii and iii we introduced the core algorithms and techniques presented in this dissertation. In this part we present how such approaches can be used for additional challenges, such as the one described in this chapter, multi-label topic classification.

Twitter is a popular service where millions of brief and instant messages about products, services or events are published per day, as we remarked in previous chapters. The vast amount of opinions and reviews provided in this micro-blogging social network is helpful in order to make interesting findings about a given industry. The Twitter search functionality can be useful to find messages about a particular product, but it is not a practical tool when we want to poll messages dealing with a given set of general topics. In this respect, this chapter presents an approach to classify Twitter messages into various topics. We tackle the problem from a linguistic angle, applying the approach described in Chapter 6, initially intended for sentiment analysis. Their practical utility has been supported by the results obtained at the TASS competition (Villena-Román et al., 2013), where an initial implementation of our approach achieved the first place in the topic classification task. We carry out a wide range of experiments to determine which kinds of linguistic information have the greatest impact on this success and how they should be combined in order to obtain the best-performing system.

8.1 DESCRIPTION

Twitter is a great meeting point where users can share their views about politics, events, technology, films and many other topics. The task of analyzing and comprehending all this information is becoming a need for companies in order to know directly from the source what is being said about them and their industry. For this purpose, they often rely on opinion mining applications for making better decisions, identifying key thoughts about their area of influence and even predicting their performance in the stock market (Li and Li, 2013; Montoyo, Martínez-Barco, and Balahur, 2012; Yu, Duan, and Cao, 2013). One of the main issues is that many of the messages under analysis are not useful for the task because they deal with unrelated topics. This may not be a serious issue in specialized forums, but it becomes a real problem when monitoring media such as Twitter, where users publish comments about all kinds of topics. In this context, applying filtering steps is necessary to be able to exploit the

messages in this social network, discriminating unrelated opinions and reducing the amount of traffic to analyze. For example, for a firm in the motion picture industry wishing to retrieve Twitter messages about a given movie, a search based on the film title can give as a result a lot of irrelevant messages in which the words appearing in the title are used in contexts unrelated to the movie domain (Rui, Liu, and Whinston, 2013).

Moreover, opinions can involve different topics, so traditional single-label classification systems are not appropriate to correctly deal with topic categorization, as users often tend to relate different subjects in the same message. The following lines illustrate some real examples[1] of tweets occurring at the corpus described in §8.2.1, that are annotated with their topics and where those are linked in different ways:

- *'The key to the new government: its structure. Will there be two deputy prime ministerships or not. The key, in the economic team'*: This tweet explicitly relates two close topics: politics and economy.
- *'The intelligent public is on social networks. Education determines their use more than wealth. Impact on media'*: This message contains information about technology, referring to social media, which also often represent a way of entertainment. Finally, the tweet was also annotated with the economy label due to making a reference to the concept of wealth.
- *'Hii my tweeps! A wonderful day to do a twitcam as I promised it will be at 6:oo pm (Mexico time) see you soon!!!'*: The tweet was assigned to the music and *other* topics. Although *a priori* it has nothing to do with music, we must take into account this tweet was addressed to his Mexican fans by the Spanish artist Alejandro Sanz (@alejandrosanz).

In order to address these issues, we propose a multi-topic classification approach for Twitter messages. We rely on linguistic information, managing lexical, syntactic, psychological and semantic knowledge by means of a NLP pipeline that includes preprocessing, tagging and parsing steps. Linguistic processing of Twitter messages is particularly challenging, as they are characterized by the use of a very informal language combined with specific Twitter elements (e. g. user mentions or hashtags). Therefore, well-performing techniques for lexical and syntactic processing of regular texts do not behave as well when confronted with Twitter messages, needing an adaptation to this new kind of text genre.

Classifier

To be able to assign several topics to the same tweet, we carried out a *one vs all* strategy: given n topics, this perspective proposes to train n

1 The original tweets were written in Spanish, but for clarity we have translated them to English.

classifiers where each one makes it possible to differentiate a topic *i*, where $i \in [1, n]$, from the others. Thus, if we plan to discover if a text is talking about a topic *X*, *Y* or *Z*, or several of them at the same time, we would create three classifiers: *topic X vs Other*, *topic Y vs Other* and *topic Z vs Other* to determine which subset of labels we should assign to the tweet. If our strategy *one vs all* always discards the topic under study (i. e. *X*, *Y* and *Z*), the system will assign to the tweet a default class.

As in Chapter 6, we relied on an SMO following the default configuration included in Hall et al. (2009) data mining software. To feed the model, we relied on the set of features already used in Chapters 6 and 7, and also consider the exploration of features with IG > 0.

8.2 EXPERIMENTS

8.2.1 *Dataset*

In this section, we are relying on the TASS corpus, a collection of tweets which has been specifically annotated to perform text analytics tasks. The corpus is composed of a training set and a test set which contain 7 219 and 60 798 tweets, respectively. In addition to the sentiment annotations, that we used in previous chapters, the corpus also contains topic labels. In particular, each tweet is annotated with one or more topics, which involve up to 10 categories: *films*, *soccer*, *economics*, *entertainment*, *literature*, *music*, *politics*, *sports* (other than soccer), *technology* and *other*. We take the *other* class as the default class.

The gold standard has been generated by a pooling of the submissions, followed by a human review by TASS organization for the thousands of ambiguous cases. Appendices B and C show the topic distribution of tweets in the collection, for both training and test sets. The classes of the training set are unbalanced. This may be interesting from a real-world environment and industry point of view, since some topics are often more popular than others, and therefore it may be difficult to build a balanced training set. In this situation, from a performance perspective, supervised methods tend to present biases when there are large differences in the number of training samples for each class. Thus, we decided to apply oversampling to the minority categories.

8.2.2 *Evaluation*

Evaluation metrics

We evaluate our approaches by means of the standard metrics for multi-label classification: Hamming loss (HL), label-based accuracy

(LBA) and exact match (EM). They are calculated according to Equations 9, 10 and 11, where:

- L is the set of labels.
- D is the set of instances of the collection.
- Y_i is the set of the labels expected for an instance i.
- Z_i is the set of labels predicted for an instance i.
- $\triangle$ is the symmetric difference operation between sets.
- tp_i is the true positive classifications for class i, where a result is a true positive iff all the gold labels and only the gold labels are assigned to the target instance.
- n is the total number of classes.

$$\text{Hamming loss} = \frac{1}{|D|}\sum_{i=1}^{|D|}\frac{|Y_i \triangle Z_i|}{|L|} \tag{9}$$

$$\text{Label–based accuracy} = \frac{1}{|D|}\sum_{i=1}^{|D|}\frac{|Y_i \cap Z_i|}{|Y_i \cup Z_i|} \tag{10}$$

$$\text{Exact match} = \frac{\sum_{i=0}^{n} tp_i}{\sum_{i=0}^{n} tp_i + \sum_{i=0}^{n} tn_i + \sum_{i=0}^{n} fp_i + \sum_{i=0}^{n} fn_i} \tag{11}$$

These metrics reflect different aspects whose relevance should depend on the type of application. We will illustrate the behavior of these metrics in Example 22:

EXAMPLE 22 (Behavior of different metrics for multi-label topic classification). Suppose two tweets, t1 and t2, where:

- $t1_a = \{politics, economy\}$, represents the actual topics for t1.
- $t1_p = \{sports, economy\}$ indicates the predicted topics for t1.
- $t2_a = \{sports, films, entertainment, football, economy\}$ refers to the actual topics for t2.
- $t2_p = \{politics, films, entertainment, football, economy\}$ represents the predicted topics for t2.

Hamming loss is a loss function, thus its optimal value is zero. It measures the number of wrong labels with respect to the total number of labels, but does not appropriately reflect the percentage of the correctly predicted labels.
Calculating the Hamming loss for t1 and t2, we obtain:

$$HL_{t1} = \frac{|t1_a \triangle t1_p|}{|L|} = \frac{|\{sports, politics\}|}{|L|} = \frac{2}{|L|}$$

$$HL_{t2} = \frac{|t2_a \triangle t2_p|}{|L|} = \frac{|\{sports, politics\}|}{|L|} = \frac{2}{|L|}$$

and so $HL_{t1} = HL_{t2}$, although t2 has a larger percentage of successful predicted topics.

Label-based accuracy is a measure able to harmonize the number of not assigned topics with respect to the wrongly selected ones. Taking again t1 and t2 as example, the LBA for each one would be:

$$LBA_{t1} = \frac{|t1_a \cap t1_p|}{|t1_a \cup t1_p|} = \frac{|\{economy\}|}{|\{politics,sports,economy\}|} = \frac{1}{3}$$

$$LBA_{t2} = \frac{|t2_a \cap t2_p|}{|t2_a \cup t2_p|} = \frac{|\{films,\ entertainment,\ football,\ economy\}|}{|\{sports,\ politics,\ films,\ entertainment,\ football,\ economy\}|} = \frac{2}{3}$$

concluding that the LBA for t2 is better than for t1.

Finally, a special case of LBA is the exact match metric, which is a more restrictive metric due to the fact that it only considers a multi-label classification as successful when $Y_i = Z_i$, that is $Y_i \cap Z_i = Y_i \cup Z_i$. If we calculate the exact match for t1 and t2 we would obtain in both cases a value of 0. Note that taking an example i where $Y_i = Z_i$ the EM, LBA and HL would be 1, 1 and 0, respectively.

Additionally, we will also take two additional metrics, used by TASS organizers, calculated according to Equations 12 and 13: *At least one* takes as valid a classification whenever at least one topic is right, whereas *Match all* considers a multi-label classification valid when a superset of the actual topic set has been predicted.

$$\begin{aligned} \text{At least one} &= \frac{1}{|D|}\sum_{i=1}^{|D|} f(i) \\ \text{where } f(i) &= \begin{cases} 1 & \text{if } Y_i \cap Z_i \neq \varnothing \\ 0 & \text{if } Y_i \cap Z_i = \varnothing \end{cases} \end{aligned} \tag{12}$$

$$\begin{aligned} \text{Match all} &= \frac{1}{|D|}\sum_{i=1}^{|D|} g(i) \\ \text{where } g(i) &= \begin{cases} 1 & \text{if } Y_i \subseteq Z_i \\ 0 & \text{if } Y_i \nsubseteq Z_i \end{cases} \end{aligned} \tag{13}$$

The drawback of these measures is that you can obtain a *perfect result* assigning all possible categories to each tweet. Therefore, they are less robust with respect to academic misconduct.

□

Results

Table 47 shows the performance for some of the basic features already used in previous chapters. The information is ordered following the exact match metric, in descending order. Bi-grams of lemmas obtained the best exact match, since the baseline (composed of only words) obtained the best Hamming loss and label-based accuracy

values. For uni-grams of words and lemmas we also included results without considering the IG of the features, in order to show the need for this step. The use of n-grams, both of words and lemmas, clearly outperforms features based on part-of-speech and psychological knowledge, which are not helpful by themselves. It is important to remark that uni-grams of words improve both the label-based accuracy and exact match over uni-grams of lemmas. But the trend is not present when using bi-grams, where lemmas perform better. We hypothesize that this is due to sparsity problems: words are sparser than lemmas; this may not be a major problem when training with n-grams where n=1, but when employing a larger value of n, combinations highly increase the dimensional space of features and probably a large set would be needed to even out the performance between bi-grams of words and lemmas.

Model	ig	hl	lba	em
Bi-grams of lemmas (BL)	0	0.077	0.626	**0.530**
Words (W) (baseline)	0	0.073	**0.658**	0.527
Bi-grams of words (BW)	0	0.080	0.613	0.524
Words (W)	-	0.079	0.634	0.498
Lemmas (L)	0	0.078	0.640	0.493
Lemmas (L)	-	**0.085**	0.611	0.460
Fine PoS-tags (FT)	0	0.289	0.262	0.032
Psychometric (P)	0	0.301	0.250	0.026
Coarse PoS-tags (CT)	0	0.384	0.186	0.003

Table 47: Performance for basic features models on the TASS topic classification corpus. The '-' character indicates that no information gain threshold was considered.

Table 48 illustrates how by combining various sets of the features proposed above we can obtain an even better exact match value. The results in Table 48 show that psychometric properties, although not being helpful by themselves when used in isolation, allow us to improve the exact match of other models based on n-grams. In this way, LIWC psychological dictionaries seem to be able to provide additional information that a model based on terms cannot represent. With respect to the label-based accuracy and the Hamming loss metrics, no combined model was able to outperform the baseline. Thus, this table outlines the performance that can be achieved by a multi-topic model based purely on lexical information and without taking meta data into account.

Finally, we also included generalized dependency triplet features in order to find out if syntactic information helps us build more ac-

curate topic classification models. Tables 49 and 50 illustrate how syntactic knowledge can improve both the exact match and the label-based accuracy over the best performing lexical-based systems. Table 51 breaks down the results by categories comparing the best model with respect to the bag-of-words approach. In particular, we obtain the best results using the following generalized dependency triplets: $\text{CT} \xrightarrow{d} \text{L}$ and $\text{L} \xrightarrow{d} \text{L}$.

After including syntactic information we finally obtain a model that achieves the best performance in the three standard metrics for multi-label classification tasks. The model is presented in Table 50 and uses *uni-grams of words*, *(coarse PoS-tag, dependency, lemma)* and *(lemma, dependency, lemma)* features to feed the SMO classifier. This suggests that syntactic information relates better with words than with n-grams of terms. We hypothesize this is because features such as bi-grams provide some structural information, which may cause some redundancy with dependency triplets. In this way, we conclude that instead of using informative features by themselves, it is more relevant to employ features that are as independent as possible from each other.

Model	**hl**	**lba**	**EM**
BL ∪ P	0.076	0.632	**0.539**
BL	0.077	0.626	0.530
W ∪ BW ∪ P	0.078	0.647	0.530
W ∪ BW	0.074	0.646	0.529
W ∪ P ∪ FT ∪ DT	**0.073**	0.655	0.527
W ∪ P ∪ FT	**0.073**	0.656	0.528
W	**0.073**	**0.658**	0.527
W ∪ P	**0.073**	0.655	0.526
BL ∪ P ∪ T ∪ DT	0.081	0.615	0.498
BL ∪ P ∪ FT	0.082	0.612	0.495

Table 48: Performance on combining lexical, syntactic, psychometric and semantic knowledge on the TASS topic classification corpus

8.3 CONCLUSION

This chapter presented a supervised topic classification system for Spanish tweets based on a linguistic perspective. We address the problem as a multi-label classification task, since a text can refer and relate several topics. We proposed an approach inspired on the features described in Chapter 6 which does not take into account any type of meta data, simply considering the information provided by the text itself.

Model	HL	lba	em
BL ∪ P	**0.076**	0.632	0.539
BL ∪ P ∪ CT $\xrightarrow{d}$ L	0.071	**0.653**	0.557
BL ∪ P ∪ CT $\xrightarrow{d}$ L ∪ L $\xrightarrow{d}$ L	0.072	0.65	**0.559**

Table 49: Performance on improving the best model, according to the EM metric, by means of generalized dependency triplets, on the TASS topic classification corpus

Model	hl	lba	em
W	**0.073**	0.658	0.527
W ∪ CT $\xrightarrow{d}$ L	0.071	0.661	0.548
W ∪ CT $\xrightarrow{d}$ L ∪ L $\xrightarrow{d}$ L	0.068	**0.669**	**0.572**

Table 50: Performance on improving the best model, according to the LBA metric, by means of generalized dependency triplets, on the TASS topic classification corpus

The approach has been applied on Twitter, given the present success of this medium, but it would be easily adaptable to other social networks, blogs or forums. The practical utility of this approach has been tested at the TASS evaluation campaign, where an initial model following this same angle was the best performing system in the topic classification task. Our experimental results provided an exhaustive evaluation through several sets of features, showing how lexical, syntactic, psychological and semantic attributes allow to improve different aspects that a topic classification system should take into account. The results lead us to conclude that relating features by means of dependency parsing adds complementary information over pure lexical models, making it possible to outperform those on standard metrics for multi-label classification tasks.

Category	f1		p		r	
	Best	**Words**	**Best**	**Words**	**Best**	**Words**
films	**0.331**	0.306	**0.331**	0.216	0.411	**0.523**
politics	0.713	**0.733**	0.747	**0.754**	0.683	**0.714**
technology	0.323	**0.344**	**0.398**	0.252	0.272	**0.540**
entertainment	0.441	**0.442**	**0.443**	0.335	0.439	**0.650**
sports	0.261	**0.271**	**0.333**	0.224	0.215	**0.341**
other	0.675	**0.689**	0.578	**0.611**	**0.811**	0.790
economy	**0.436**	0.391	**0.359**	0.267	0.556	**0.729**
music	**0.445**	0.436	**0.445**	0.436	0.594	**0.710**
football	0.297	**0.332**	**0.481**	0.301	0.215	**0.371**
literature	**0.380**	0.348	**0.395**	0.255	0.366	**0.548**
Macro average	**0.430**	0.429	**0.452**	0.353	0.450	**0.590**

Table 51: Detailed performance per categories both for the best syntactic model and the bag-of-words approach. If a tweet discusses *films* and *entertainment*, but it is only classified in the *films* class, it would be taken as a true positive for the *films* category, and as a false negative for the *entertainment* class

9

POLITICAL ANALYSIS

In this chapter we show how Spanish SentiStrength (Chapter 3) can be applied to analyze political tendencies on political tweets in real-time. The aim is not to predict elections, but to assess whether Twitter can reveal changing perceptions about politicians over time and the influence of individual events. We describe existing related work to then explain how we collected and cleaned up the data, computed overall scores for different parties and politicians and compared them against results provided by traditional polls.

9.1 DESCRIPTION

The online component of politics is widely recognized as important. Barack Obama is sometimes cited as the first major politician to effectively harness web networks for traditional political purposes (Harfoush, 2009) and Facebook, Twitter and YouTube played an important role in the Arabic Spring movement (Howard et al., 2011). In Spain, the success of the popular 15M political protest movement was partly due to organizing through social media (Borge-Holthoefer et al., 2011). It is therefore important to develop methods to analyze social media to gain insights into the online components of political participation. Although the traditional way to measure popularity or voting intentions is through opinion polls (CIS, 2014), these, even when reliable, are costly and time-consuming. This is a particular problem for regional or local elections, where exhaustive surveys are infeasible given the number of districts and candidates. In addition, polls are usually published every few months, so it is not possible to find out the impact that an individual act or decision has on society. As a partial solution, social media analyses may be able to track offline opinions through their online reflections.

As briefly discussed below, there are many different automatic methods to extract opinions and trends from social media. In particular, the field of SA made it possible to automatically detect opinions on a large scale. Whilst some studies have applied these methods to political topics, most research has focused on consumer reviews of products and most systems are designed exclusively for English text. There is therefore a need for political sentiment analysis systems as well as political sentiment analyses for languages other than English.

In response to the above gap, this chapter applies Spanish SentiStrength, the sentiment analysis system presented in Chapter 3, to Spanish political tweets with a case study of the main Spanish po-

litical representatives and parties. The results are based on a set of 2 704 523 tweets mentioning 30 politicians and 6 political parties over 41 days. All the analyses in this chapter were conducted before the polling results were published and can therefore be considered to be real predictions.

9.2 POLITICS IN SOCIAL MEDIA

Some research has analyzed the use of social media by politicians. For example, a Dutch study found that national election results correlated with politicians' use of social media, but the same was not true for local elections (Effing, Hillegersberg, and Huibers, 2011). In contrast, an analysis of Australian politicians' use of Twitter argues that it is difficult to control, interpret or understand the benefit that they gain from it (Grant, Moon, and Grant, 2010). However, many of the literature discusses tweeting about politics by the electorate rather than by politicians.

Most political analyses on Twitter have focused on predicting electoral outcomes (Ceron, Curini, and Iacus, 2015b), but Twitter can also be used to identify political preferences (Golbeck and Hansen, 2011) and for day-to-day monitoring of electoral campaigns (Ceron, Curini, and Iacus, 2015a; Jensen and Anstead, 2013; Wang et al., 2012). One of the first studies analyzed 104 003 Twitter messages mentioning the main German parties or politicians before the 2009 federal elections (Tumasjan et al., 2010). The German tweets were automatically translated into English for a LIWC keyword analysis (Pennebaker, Francis, and Booth, 2001). The numbers of tweets mentioning parties or politicians were found to closely reflect voter preferences in traditional election polls. This study showed that Twitter may complement traditional polls as a political forecasting tool. Nevertheless, a Twitter sample may not be representative of the electorate, the general sentiment dictionaries used may not be optimal for politics, and replies to political messages may not be captured by keyword searches (Tumasjan et al., 2010). In support of the latter point, keyword-based searches for political tweets can aim for high precision (i. e. they generate few false matches) but not high recall (i. e. they will miss many relevant tweets) (Marchetti-Bowick and Chambers, 2012).

A time series analysis of the 2008 US presidential elections derived day-to-day sentiment scores by counting positive and negative messages: a message was defined as positive if it contained a positive word, and negative if it contained a negative word (a message can be both positive and negative). Although there were many falsely detected sentiments, these errors may tend to cancel out (O'Connor et al., 2010). This approach missed sentiments in tweets using nonstandard spellings and emoticons and needed smoothing to stabilize the results. The sentiment results correlated with presidential ap-

proval polls, but not with election polls, and message volume did not have a straightforward relationship with public opinion. Another study of US elections also found that sentiment results did not predict election outcomes, possibly due to the overrepresentation of young people and Democrats in Twitter (Gayo-Avello, 2011). Almost identically for the 2011 Irish General Election, sentiment did not predict voting patterns due to the overrepresentation of one party and the underrepresentation of another, although a simple volume measure was more accurate than sentiment (Bermingham and Smeaton, 2011). Twitter bigrams (consecutive words) can also be used to predict daily approval ratings for US presidential candidates using a time series regression approach (Contractor and Faruquie, 2013).

Election results have also been predicted in Twitter for many other countries, with varying degrees of success. For the 2013 general election in Italy tweet volume was a reasonable indicator for the final results, detecting a strong presence in Twitter of the (unexpected) winning party and the (also unexpected) relative weakness of another party, but failing to make accurate predictions for small parties, who were overrepresented in Twitter (Caldarelli et al., 2014). Small party overrepresentation in Twitter has also been found for German elections (Jungherr, Jürgens, and Schoen, 2012). Tweet volumes were a reasonable indicator for the 2011 Nigerian Presidential election (Fink et al., 2013) and the Venezuelan, Paraguayan and Ecuadorian Presidential elections of 2013 – especially when counting tweets mentioning the full names of candidates or mentioning the aliases of candidates jointly with an electoral keyword Gaurav et al., 2013. In contrast, Twitter did not seem to be able to predict the 2011 Dutch Senate election outcomes (Sang and Bos, 2012).

One particularly comprehensive study analyzed 542 969 tweets mentioning candidates together with data on 795 election outcomes in 2010 and 2012 US elections and socio-demographic and control variables such as incumbency, district partisanship, median age, percentage white, percent college educated, median household income, percentage female and media coverage (DiGrazia et al., 2013). There was a statistically significant association between the number of tweets mentioning a candidate and their subsequent electoral performance. The models under-performed in relatively uncompetitive or idiosyncratic districts, however.

Despite some of the positive results reported above, electoral predictions on Twitter data overall tend not to be better than chance (Metaxas, Mustafaraj, and Gayo-Avello, 2011). When the predictions are better than chance (e. g. Gayo-Avello, Metaxas, and Mustafaraj (2011)), they are not an improvement on simply predicting that all incumbents would be re-elected (see also: Huberty (2013)). It follows that sentiment analyses need to be sophisticated in order to make credible election predictions (Gayo-Avello, 2012).

Several studies have applied sentiment analysis to social web politics and used the results to identify patterns of behavior rather than to predict elections. It is possible to predict how Twitter users will vote by comparing the language of their tweets with that of the main parties in an election (Makazhanov and Rafiei, 2013) and this technique has been used to show, unsurprisingly, that politically active users are less prone to changing their preferences. It is also possible to estimate the level of disaffection across society by counting negative tweets about politics in general, and this approach has shown that peaks in disaffection can correlate with important political news (Monti et al., 2013). Twitter has also been used to study divisions within electorates. A sentiment analysis of Twitter in Pakistan, for example, found differences between expatriates and people living in the country and between urban and rural areas (Razzaq, Qamar, and Bilal, 2014).

Finally, Twitter is also used by journalists to add direct quotes from politicians to stories, and so Twitter sometimes helps to generate the news in addition to reflecting it (Broersma and Graham, 2012).

9.2.1 *Twitter as a tool for political analysis in Spain*

Twitter is extensively used in Spain for politics during elections (Criado, Martínez-Fuentes, and Silván, 2013). An analysis of 370 000 tweets from over 100 000 users during the 2011 Spanish general elections found that half of the messages were posted by 7% of the participants, 1% of users were the target of half of the mentions, 78% of the mentions were of politicians, 2% of the users caused half of the retweets and the source of 63% of the retweeted messages were mass media accounts (Borondo et al., 2012). Moreover, 65% of the participants were men, with Madrid overrepresented but no significant differences were found between the behavior of those living in large cities and in the rest of Spain; citizens with a strong party identification were particularly active (Barberá and Rivero, 2012).

A study of 84 387 tweets from Catalan regional elections found Twitter users to cluster by political affinity (Congosto, Fernández, and Moro, 2011), corroborating similar results from other countries (Conover et al., 2012; Conover et al., 2011; Livne et al., 2011). Despite this, it is difficult to predict the party of a Twitter user from the list of accounts that they follow (Barberá, 2012). Ideological groupings also occur on the web for political and media websites in Spain, highlighting the partisan nature of the media (Romero-Frías and Vaughan, 2012).

The number of times that Spanish political parties are mentioned on Twitter seems to correlate with electoral outcomes, but only for parties that obtained more than 1% of votes (Barberá and Rivero, 2012). One study focused on predicting results for a new small, Span-

ish green party, eQuo, with an electoral strategy based mainly on social media (Deltell, 2012). For several days its proposals were trending topics on Twitter, and its Facebook page was more visited and had more "likes" than the pages of the other political parties. Nevertheless, this successful social media campaign did not translate into any elected politicians. Perhaps surprisingly, eQuo performed best in districts in which it used traditional activities, such as meetings and posters.

Twitter has been used for successful predictions for the Andalusian regional elections of 2012, counting the followers of the Twitter accounts of political parties and their leaders. For the two major parties, Partido Popular and Partido Socialista Obrero Español, this simple method gave results that were closer to the final election outcomes than were traditional polls (Deltell, Claes, and Osteso, 2013), although the polls were particularly inaccurate in these elections. The Twitter follower method was inaccurate for small and new parties, including those, such as the Izquierda Unida, with leaders that were inactive on Twitter.

In an academic competition to classify the political tendency of public figures (not necessarily politicians) into left, centre, right or neutral (Villena-Román et al., 2013), the best performing system considered a number of politicians related with the main political parties (Pla and Hurtado, 2013). If messages from a user contained one of these politicians tended to be negative then the user was classified against that political orientation, and vice versa. Another competition was to classify the polarity of tweets mentioning one of the four main national parties. The best performing system assumed that the polarity of the whole tweet corresponded to the polarity of the party (Gamallo, García, and Fernández Lanza, 2013).

9.3 MATERIALS AND METHODS

We now proceed to describe our approach to estimate the point of view of the Spanish society with presence in Twitter with respect to main politicians and political parties of Spain. The analysis was carried out between December of 2014 and January of 2015.

For each of the six main political parties in 2014, we selected five important politicians (see Appendix A). Since some parties have few widely recognized members (CIS, 2014), we took into account the number of Twitter followers to ensure that the selected politicians could be discussed on Twitter:

- *Partido Popular* (PP): The main conservative party and winner of the 2011 elections. Its leader and prime minister is Mariano Rajoy.
- *Partido Socialista Obrero Español* (PSOE): The main social-democratic party and in government until 2011. Its secretary-general until 2016 was Pedro Sánchez.

- *Izquierda Unida* (IU): A left-wing party and usually third in general elections. Its leader from that time, Cayo Lara, (@cayolara) was set to step down and be replaced by Alberto Garzón.
- *Unión, Progreso y Democracia* (UP YD): A political party founded in 2007. Its leader was Rosa Díez until 2015, and during the polling period it was the only main politician without an official Twitter account. The parlamentarial presence of the party dissapeared after the general elections of 2015.
- *Ciudadanos (*CS*)*: A non-regionalist centre party originally from Catalonia and led by Albert Rivera .
- *Podemo*s: A new left wing political party born in January 2014. The elected leader is Pablo Iglesias and at least one poll has rated them as the most popular Spanish party (CIS, 2014) at the moment this study was carried out.

We collected tweets from 3 December 2014 to 12 January 2015 via the Twitter Streaming API. A number of steps were taken to filter out irrelevant tweets:

- Tweets just containing information without an opinion (neutral tweets that received neither a positive and a negative strength), were removed.
- Retweets of tweets from the parties or politicians analyzed were removed. This step was taken because these messages tend to be retweeted many times due to their number of followers and the author of the message and therefore seem to create false peaks in activity.
- Messages involving two or more different political parties were removed.
- Phrases quoted in tweets were removed because these are often associated with titles or rhetorical devices, such as sarcasm or irony, that should be treated differently (Thelwall, Buckley, and Paltoglou, 2012).

After the above filtering steps we obtained a total dataset of 2 704 523 tweets.

9.4 EXPERIMENTS

We first computed daily average positive and negative sentiment scores for each politician and party using the Spanish SentiStrength system presented in Chapter 3. Polls from the reputable Centro de Investigaciones Sociológicas (CIS, www.cis.es) (CIS, 2014) were taken as the primary source of public opinion in Spain and used as the reference point for the Twitter results. In order to cover a wider set of entities, additional well-known polls carried out by private companies were

also included: Invymark[1], GESOP[2], DYM[3], Sigma-2[4] and Termómetro electoral [5]. The CIS poll used covered January 4 to January 12, 2015.

Table 52 compares the sentiment for the political leaders with the national poll results. Polls differ in their criteria and coverage, and the CIS poll includes only four leaders. The dual positive and negative *sentistrength* scores make it possible to assess whether the most hated leaders are also the most loved, and whether some politicians attract particularly strong emotions. The ranking provided by *sentistrength* for positivity matches exactly that provided by CIS. Surprisingly, *sentistrength*'s negativity rank is also similar to the one provided by CIS, switching only third and fourth place. The similarity between rankings can be compared with Hamming loss distance (Equation 14) and the out-of-place measure from Cavnar and Trenkle (1994) (Equation 15). Table 53 shows the results. For example, given a ranking, R, where the system only failed the classification for the entity E: predicted(E) = 5 and gold(E) = 2, the score for the Hamming-loss(R) would be 2, since we need to make two changes to obtain the correct ranking, while the out-of-place(R) would be 3, because that was difference between the predicted and gold position. However, if predicted(E) is 3 then out-of-place(R) would be 1, although the Hamming loss would be the same.

$$\mathrm{Hamming\ loss} = \sum_{i=1}^{n} f(\mathrm{predicted}(i), \mathrm{gold}(i)) \tag{14}$$

where:

$$f(x, y) = \begin{cases} 1 & x \neq y \\ 0 & x = y \end{cases}$$

$$\mathrm{Outofplace} = \sum_{i=1}^{n} |\mathrm{predicted}(i) - \mathrm{gold}(i)| \tag{15}$$

The *sentistrength* scores were used to create daily rankings, and Mann-Whitney U tests (McKnight and Najab, 2010) were used to compare the relative rankings for pairs of politicians, as shown in Table 54. A significant result is evidence that the daily *sentistrength* averages consistently indicate one of the two politicians as being the most popular, with a better than random chance. Surprisingly, the leaders of the first and second ranked parties (CIS, 2014) were the only ones with stronger negative sentiment than positive sentiment. This suggests that negative expressions in Twitter do not imply less electoral support but could reflect other factors, such as the need for other parties to attack the leading contenders.

1 www.invimark.es
2 www.gesop.net
3 www.gesop.net
4 www.sigmados.com
5 termometroelectoral.blogspot.com.es

Table 52 also compares our ranking with the ones provided by the other polls collected. Table 53 measures how similar these rankings are to the ones provided by the main national polls, except for Sigma-2. Some of the polls were not published in January, so we substituted the ones released in December, which also fell within our period of analysis. Positive perception is a better indicator than negative perception in most of the cases to predict rankings similar to those of the traditional polls. The comparison shows that centre-right leaders (Rivera, Díez and Rajoy) are equivalently located in all polls (except Sigma-2). The differences in rankings are due to left-wing leaders (Sánchez, Garzón and Iglesias), an issue also observed when comparing traditional polls between them.

Leader	pos strength	neg strength	Cis (Jan. 2014)	Term	Invy	gesop	Sigma-2	dym
Albert Rivera	2.57	1.80	-	4.73	4.18	4.62	3.76	4.20
Pedro Sánchez	2.30	1.88	3.68	4.02	4.18	4.56	3.81	3.71
Alberto Garzón	2.25	2.03	-	4.52	3.89	4.60	3.80	3.90
Pablo Iglesias	2.22	2.23	-	4.51	3.96	4.51	3.93	4.00
Rosa Diéz	2.19	2.16	3.66	3.02	3.54	4.15	3.87	3.70
Cayo Lara	2.13	2.02	3.53	-	-	-	-	-
Mariano Rajoy	2.07	2.32	2.24	2.81	3.27	2.65	3.43	2.60

Table 52: Average *sentistrength* scores vs. national Spanish poll scores. Reference polls ratings range from 0 to 10.

Metric	*sentistrength*	Cis (Jan. 2014)	Term	Invy	gesop	Sigma-2	dym
Hamming-loss distance	POS	0	3	3	2	5	2
	NEG	2	4	2	3	5	3
Out-of-place measure	POS	0	4	2	2	12	4
	NEG	2	6	4	4	12	6

Table 53: Predicted and gold standard rankings compared to Hamming-loss distance and out-of-place measure.

With respect to political parties, it does not make sense to predict poll results with sentiment because the two are different. For example, according to a CIS (2015) poll, Mariano Rajoy and the Partido

Politicians	Class	Albert Rivera	Pedro Sánchez
Albert Rivera	POS	-	-
	NEG	-	-
Pedro Sánchez	POS	0.0026	-
	NEG	0.0500	-
Alberto Garzón	POS	0.0011*	0.3050
	NEG	0.0091	0.0347
Pablo Iglesias	POS	0.0004*	0.1069
	NEG	0.0000*	0.0000*
Rosa Díez	POS	0.0023*	0.1149
	NEG	0.0002*	0.0015*
Cayo Lara	POS	0.0000*	0.0170*
	NEG	0.05843*	0.1360*
Mariano Rajoy	POS	0.0000*	0.0000
	NEG	0.0000*	0.0000*

Table 54: Mann-Whitney U test, at a confidence level of 95%. ($p < 0.05$) for Albert Rivera and Pedro Sánchez against the main Spanish political leaders. This shows the complete results for the best two scored leaders (Albert Rivera and Pedro Sánchez), although Bonferroni corrections were applied to counteract the problem of multiple comparisons taking into account all possible comparisons between the pairs of leaders ($p < 0.002381$ to accept that differences in perception are significant). Cells marked with '*' indicate significant differences.

Popular are the least popular leader and party, but would get the most votes. As shown in Table 55, sentiment scores are not reliable for predicting elections. The number of tweets naming either the political party or their leader is a better indicator, confirming similar results for other countries (Contractor and Faruquie, 2013; Tumasjan et al., 2010).

In general, the more conservative the party is, the more negative tweets mention it and this may reflect a bias in the user base of Twitter, such as towards young people. Younger people may tend to be left-wing (Pew-Research-Center, 2011), which would explain the online hostility to the right. Left wing young people may also be more politically active (Oswald and Schmid, 1998), exacerbating the bias. In Spain, according to the CIS poll 40% of the population are left-wing and 21% are right-wing, and so negativity towards the right could be expected even without the youth bias.

	@Ciudadanoscs	@ahoraPodemos	@UP yD	@PSOE	@iunida	@Ppopular
POS strength	2.59	2.22	2.07	2.11	2.04	2.05
NEG stregtnh	-1.63	-2.05	-2.22	-2.13	-2.15	-2.48
Left- (0) Right-wing (10)	5.14	2.28	5.35	4.62	2.62	8.17
Vote intention +sympathy	1.9	19.3	2.5	18.1	4.2	14.6
Vote estimate	3.1	23.9	4.6	22.2	5.2	27.2
Average daily party mentions	963	19 495	2 218	4 881	542	2 858
Average daily leader mentions	893	4 353	289	2957	623	14 007

Table 55: Average positive and negative *sentistrength* in tweets mentioning the main Spanish political parties.

Table 56 shows the rankings for all politicians. Traditional polls do not provide surveys for many of these, and so the ranking cannot be compared with other rankings. Nevertheless, they give information that cannot be obtained from traditional polls. The Partido Popular politicians are last in both rankings, reinforcing the online sentiment agreement with traditional polls. Similarly, Ciudadanos politicians had the highest scores, which reflects the results for their leader, Albert Rivera. These results also show that politicians coming from these two parties attract similar sentiments to their party overall. The same is true for PSOE politicians, except that the party account is an outlier. Politicians from Podemos, IU and UP yD were more scattered in rakings, but these might reflect specific news stories with wide media coverage. The low average positive sentiment for Tania Sánchez (IU) was perhaps reflected by her resignation shortly after the period of analysis. Negative press coverage about her management of a town council and disagreements with IU in December (Riveiro, 2014) and

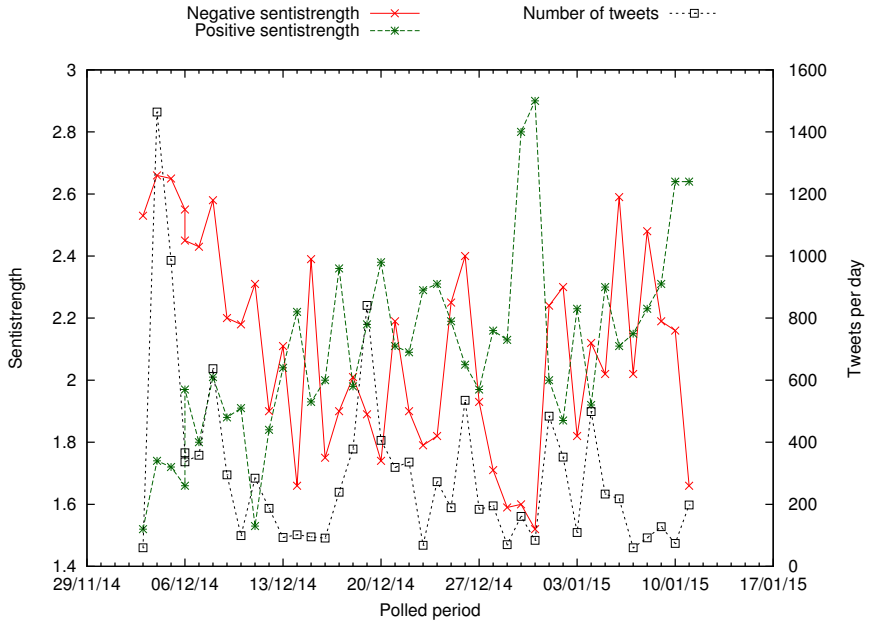


Figure 10: Variation of positive and negative perception of Iñigo Errejón during the period of polling

January (Gil, 2014; Silió, 2014) seemed to trigger her departure. Iñigo Errejón (Podemos) also had low sentiment rankings apparently as a consequence of negative press coverage. Figure 10 shows how during the beginning of the polling period he had very low scores, coinciding with news about alleged irregularities at his previous job from December 4, 2014 (Granado, 2014; RTVE.es/EFE, 2014). A peak in the number of tweets at this time confirms that Twitter sometimes reflects popular political events.

9.5 CONCLUSION

In this chapter, we collected a large amount of Spanish political tweets in a month and used the Spanish SentiStrength system presented in Chapter 3 to then analyze the sentiments expressed about the main Spanish politicians and parties. The sentiment scores obtained by SentiStrength were used to build ranks for the politicians and their parties, giving popularity ratings that are comparable with those provided by the classic polls, although tweet volume was a much better predictor of voting intentions. An advantage of sentiment analysis in Twitter is that is can be more comprehensive than traditional polls by covering more parties and politicians, although the results are less reliable for smaller parties. A deeper analysis of politicians that had sentiment scores that did not match those of their parties suggested that these had attracted negative media publicity that had been amplified in Twitter. This shows that the Twitter results may be useful to analyze the trajectories of individual politicians and perhaps even

NEG sentiment (from the lowest to the highest)			POS sentiment (from the lowest to the highest)		
Luis Salvador	CS	-1.39	Luis Salvador	CS	2.84
Fran Hervias*	CS	1.42	Fran Hervias*	CS	2.63
Inés Arrimadas*	CS	-1.59	Ciudadanos	CS	2.59
Ciudadanos	CS	-1.63	Albert Rivera	CS	2.57
Carme Chacón	PSOE	-1.80	Inés Arrimadas*	CS	2.43
Albert Rivera	CS	-1.80	Pablo Echenique	Podemos	2.31
César Luena	PSOE	-1.85	Pedro Sánchez	PSOE	2.30
Pedro Sáchez	PSOE	-1.88	Elena Valenciano*	PSOE	2.28
Susana Díaz	PSOE	-1.92	Cristina Cifuentes	PP	2.28
Elena Valenciano*	PSOE	-1.97	Carme Chacón	PSOE	2.26
Pablo Echenique	Podemos	-2.02	Alberto Garzón	IU	2.25
Cayo Lara	IU	-2.02	Irene Lozano	UP YD	2.24
Alberto Garzón	IU	-2.03	César Luena*	PSOE	2.23
Irene Lozano	UP YD	-2.03	Teresa Rodríguez	Podemos	2.22
Podemos	Podemos	-2.05	Pablo Iglesias	Podemos	2.22
Carlos Martínez Gorriaran	UP YD	-2.08	Podemos	Podemos	2.22
Teresa Rodríguez	Podemos	-2.10	Esperanza Aguirre	PP	2.20
Iñigo Errejón	Podemos	-2.10	Juan Carlos	Podemos	2.20
Toni Cantó	UP YD	-2.10	Rosa Díez	UP YD	2.20
PSOE	PSOE	-2.13	Javier Nart	CS	2.18
Gaspar Llamazares*	IU	-2.14	Carlos Martínez Gorriaran	UP YD	2.17
Izquierda Unida	IU	-2.15	Soraya Sáenz de Santamaría	PP	2.16
Juan Carlos Monedero	Podemos	-2.16	Hugo Martínez Abarca*	IU	2.16
Tania Sánchez	IU	-2.16	Toni Cantó	UP YD	2.15
Rosa Díez	UP YD	-2.16	susana Díaz	PSOE	2.15
Javier Nart	CS	-2.18	Cayo Lara	IU	2.13
Hugo Martínez	IU	-2.15	PSOE	PSOE	2.10
Cristina Cifuentes	PP	-2.20	Iñigo Errejón	Podemos	2.08
Esperanza Aguirre	PP	-2.20	Mariano Rajoy	PP	2.07
UP YD	UP YD	-2.22	UP YD	UP YD	2.07
Pablo Iglesias	Podemos	-2.23	Partido Popular	PP	2.05
Soraya Sáenz de Santamaría	PP	-2.27	Izquierda Unida	IU	2.04
Mariano Rajoy	PP	-2.32	María Dolores de Cospedal	PP	2.04
Partido Popular	PP	-2.48	Gaspar Llamazares*	PP	2
María Dolores de Cospedal	PP	-2.65	Tania Sánchez	IU	1.98

Table 56: Positive and negative sentiment ranking from SentiStrength for the tweets mentioning the politicians analyzed. '*' indicates that the politician has less than 120 mentions per day (20% of the maximum), and so may have an unreliable ranking

to evaluate the impact of negative press coverage on their popular perception.

10 ONLINE REPUTATION

Identifying how people relate aspects and traits such as performance, services or leadership with their business, is a good starting point for monitoring the perception of the public via sentiment analysis applications. In a similar line, companies are interested in user profiling: identifying the profession, cultural level, age or the level of influence of authors in an specific domain may have potential benefits when making decisions with respect to advertisement policies, for example.

In this chapter, we review our participation at RepLab 2014 (Amigó et al., 2014), a competitive evaluation for reputation monitoring on Twitter. The following tasks were proposed by the organizers: (1) categorization of tweets with respect to standard reputation dimensions and (2) characterization of Twitter profiles, which includes: (2.1) identifying the type of those profiles, such as journalist or investor, and (2.2) ranking the authors according to their level of influence on this social network. We consider an approach based on the application of natural language processing techniques in order to take into account part-of-speech, syntactic and semantic information, similar to the approach described in Chapter 6. In this chapter we describe our participation at tasks (1) and (2.2). However, each task is addressed independently, since they respond to different requirements. The official results confirm the competitiveness of our approaches, which achieved the 2nd place, tied in practice with the 1st place, at the author ranking task; and 3rd place at the reputation dimensions classification tasks.

10.1 DESCRIPTION

We below detail the tasks proposed by the organizers and how we have addressed them.

10.1.1 *Task 1: Reputation Dimensions Categorization*

The task consisted of relating tweets with the standard reputation dimensions proposed by the Reputation Institute and the RepTrak model[1]: *products&services*, *innovation*, *workplace*, *citizenship*, *governance*, *leadership*, *performance and undefined* (if a tweet is not assigned to any of

1 http://www.reputationinstitute.com/about-reputation-institute/the-reptrak-framework

the other dimensions). We addressed the task following the approach described in Chapter 6.

Dataset

The RepLab 2014 corpus is composed of English and Spanish tweets extracted from the RepLab 2013 corpus (Amigó et al., 2013), which contained a collection of tweets referring to up to 61 entities. The RepLab 2014 corpus only takes into account those who refer to banking or automotive entities, where each one is labeled with one of the standard reputation dimensions. To create the collection the canonical name of the entity was used as a query to retrieve the tweets which talk about it. Thus, each tweet contains the name of an entity. In addition, the corpus provides information about the author of each tweet, the content of external links that appear in a message and a flag to know if the tweet is written in English or Spanish.

Runs

Two different runs were sent to address the task. For each of them, we trained two different liblinear classifiers (Fan et al., 2008): one for English and another one for Spanish language. We tuned the weights for the majority classes (*products, citizenship, undefined and governance*) using a value of 0.75, giving the less frequent categories a weight of 1. The features used in each run were:

- *Run 1*: The English model took as features: uni-grams of lemmas, bi-grams of lemmas, and word psychometric properties. With respect to the Spanish classifier, the experimental setup showed that the best-performing model over Spanish messages was composed of: uni-grams of lemmas, bi-grams of lemmas and generalized triplets of the form $ø \xrightarrow{d} L$ (i. e. dependency triplets where the head is omitted). In both cases, we tried to obtain the best sets of features via greedy search on the training corpus and a 5-fold cross-validation.
- *Run 2*: This model uses the same classifier and the same sets of features as run 1, but excluding those which include the name of any of the entities used to create the training corpus. Our main aim was protecting our model from a possible bias on the training corpus. We observed that many tweets belonging to certain entities were labeled mainly only into a single reputation dimension. We were concerned that this fact could create an overfitted model which would not work properly on the test set. In this respect, this run also allowed us to measure the impact on performance of using the name of entities on the test set.

In both cases, our approaches only handle the content of a tweet, discarding the user information and the content of the external links.

In the latter case, we think processing the content of the web pages referred to in a tweet may excessively increase the cost of analyzing a tweet. In addition, we believe the tweet reputation dimensions are not necessarily to be related with the content of the link, where probably many concepts and ideas appear. The results presented below these lines seem to confirm our hypothesis since we ranked 3rd, very close to the best-performing system.

Results

Table 57 shows the performance of our systems for the reputation dimension task, based on their accuracy. For clarity reasons, we draw the performance of the best and worst systems, our runs and the baseline of the RepLab organization, a naive bag-of-words approach trained on a SVM. A detailed description of the performance of every run can be found at Amigó et al. (2013) description paper. Our run 1 ranked 3rd, confirming the effectiveness of our perspective. The second run also worked acceptably, although performance dropped by almost two percentage points. This confirms a slight bias on the test set, since it contains tweets that refer to the same entities as the training set and they were collected in the same interval of time. Table 58 show the detailed performance for our best run. Our model obtains both an acceptable recall and precision for the most prevalent classes, but the same is not true for minority classes, due to the small number of samples in the training set. The majority of the participants exhibited this same weakness.

Team	**Accuracy**
Best system	0.731
Our run 1	**0.717**
Our run 2	**0.699**
baseline-replab	0.622
Worst system	0.357

Table 57: Ranking for task 1 at RepLab 2014: Reputation Dimensions Categorization

10.1.2 *Task 2.2: Author ranking*

The task focuses on classifying authors as *influential* or *non-influential*, as well as ranking them according to that level of influence.

Category	r	p	#tweets	% tweets
Innovation	0.085	0.271	306	1.09
Citizenship	0.732	|bf 0.848	5027	17.89
Leadership	0.200	0.484	744	2.65
Workplace	0.274	0.527	1124	4.00
Governance	0.461	0.697	3395	12.08
Performance	0.404	0.499	1598	5.69
Products&Services	**0.879**	0.702	15903	56.60

Table 58: Detailed performance for our best run on the reputation dimensions categorization task

Dataset

The training and the test set are composed of the authors who wrote the automotive and banking tweets that we mentioned previously. In addition to user information, the organizers included the identifiers of the last 600 tweets of each user at the moment of the creation of the corpus. The proportion in the training set is about 30% of influential users, with the remaining 70% being non-influential.

Evaluation metrics

The organizers address the problem as a traditional ranking information problem using the *mean average precision* (MAP) (Buckley and Voorhees, 2000) as standard metric. The experimental results are ordered according to the average of automotive and banking (MAP) measures.

$$MAP = \frac{1}{tp_{influencers}} \sum_{i=1}^{N} P(i)q(i) \tag{16}$$

where:

- N is the total number of users.
- $tp_{influencers}$ is the number of true positive for the influential users.
- $P(i)$ is the precision at rank i
- $$q(i) = \begin{cases} 1 & \text{if } i^{th} \text{ user is influential} \\ 0 & otherwise \end{cases}$$

Runs

Classification of influential and non-influential users is made via a liblinear classifier, following a machine learning perspective. To rank the authors we take as the starting point the confidence factor reported by the classifier for each sample. A higher confidence should indicate a higher influence. With respect to non-influential users, we firstly negate that factor, obtaining in this way lower values for the least influential authors. We again sent two models to evaluate this task, although in this case the runs present significant differences:

- *Run 1*: A bag-of-words model which takes each word of the Twitter profile descriptions to feed the supervised classifier. The weights of the classes were tuned taking 1.8 and 1.3 for influential an non-influential users, respectively. Since the corpus is domain-dependent (automotive and banking tweets) we hypothesize that the brief biography of the user may be an acceptable indicator of influence. We observed that words such as *'car'*, *'business'* or *'magazine'* were some of the most relevant tokens in terms of information gain.
- *Run 2*: This run follows a meta-information perspective, taking the information provided by the Twitter API for any user. More specifically, we used binary features such as: URL in the Twitter profile, verified account, profile user background image, default profile, geo enabled, default profile image, notifications, translation enabled and contributors enabled. In addition the following numeric features are taken into account: *listed count*, *favorites count*, *followers count*, *statuses count*, *friends count* and *following*.

Results

Table 59 illustrates the official results for this task. The baseline of the RepLab organizers ranks the authors by their number of followers. Our run 1 achieved the 2nd place, tied in practice with the 1st place, reinforcing the validity of the proposal for a specific domain. On the other hand, our second run did not work as expected, although it outperformed the baseline.

Team	**MAP**
Best system	0.565
Our run 1	**0.563**
Our run 2	**0.403**
baseline-replab	0.378
Worst system	0.349

Table 59: Ranking for task 2.2 at RepLab 2014: Author ranking

10.2 CONCLUSION

In this chapter we addressed the challenge of online reputation (reputation classification and ranking of infuential authors) in the context of the evaluation campaign of RepLab 2014. The classification for the reputation dimensions task was addressed from an NLP perspective, including part-of-speech tagging and dependency parsing. We extracted lexical, psychometric and syntactic-based features, which were used to feed a supervised classifier. We ranked 3rd, very close to the best performing system, confirming the effectiveness of the approach. The author ranking challenge was addressed from a different perspective. We obtained the second best-performing system, tied in practice with the 1st place, by training a bag-of-words classifier which takes the Twitter profile description of the users as features. This model clearly outperformed our second run based on metadata such as the number of favorited tweets or followers.

11

SEMEVAL & SENTIMENT ANALYSIS IN TWITTER

The SemEval organization has included in recent years sentiment analysis challenges as part of its set of evaluation campaigns. In particular, in its 2016 edition the proposed subtasks related to SA in Twitter were: (1) classification into two, three and five classes and (2) quantification into two and five categories. A detailed description of them can be found in Nakov et al. (2016a). We are here describing our participation at task (1). A detailed description of our participation at subtask (2) can be found in Vilares et al. (2016).

The results and official rankings located us: 2nd (practically tied with 1st) for the binary classification task and 4th (practically tied with 3rd) for the five-class polarity classification challenge. We describe our approach below these lines.

11.1 DESCRIPTION

We address the sentiment classification subtask into two, three and five classes from a machine learning perspective. Our aim as usual is to train prediction hypothesis functions to solve classification into five (strong positive (POS+), positive (POS), neutral (NONE), negative (NEG) and strong negative (NEG+)), three (POS, NONE and NEG) and two (POS and NEG) classes. In particular, for this task we have trained a convolutional neural network using pretrained Twitter word embeddings, so that we could extract the hidden activation values from the hidden layers once some input had been fed to the network, and include them as features to feed an external classifier, an approach standardized with the name of *deep features* (Zhou et al., 2014). We describe now the process in detail below.

11.1.1 *Convolutional neural network and deep features*

As a starting point, we train a deep neural network (DNN), in particular a convolutional neural network (CNN), following a similar configuration to the one used by (Severyn and Moschitti, 2015). Figure 11 illustrates the topology of the CNN from where we will extract the hidden activation values.

Embeddings layer

Let w be a token of a vocabulary V, a word embedding is a distributed representation of that token as a low dimensional vector $v \in \mathbb{R}^n$. In

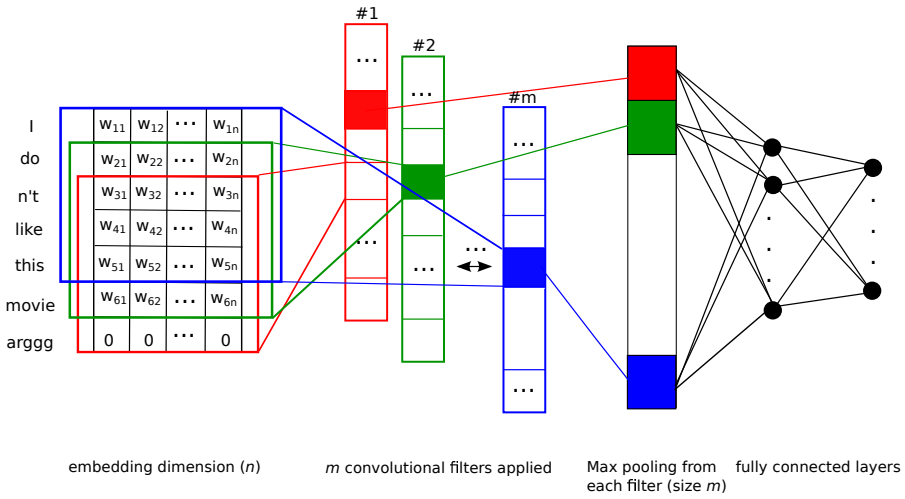


Figure 11: Topology of our CNN from where we will extract the neural activation values

that way, it is possible to create a matrix of embeddings, $E \in \mathbb{R}^{|V| \times n}$, to act as the input layer to the CNN. Particularly, we rely on a collection of Twitter word embeddings pretrained with Glove[1] (Pennington, Socher, and Manning, 2014) with $|V| \approx 10^6$ and n=100.

Thus, given a tweet t=$[w_1, w_2, ..., w_t]$, after running our input layer we will obtain a matrix $T \in \mathbb{R}^{|t| \times n}$ that will serve as the input to the convolutional layer. Since tweets might have variable length, |t| is set to 100, padding with zeros if the tweet is shorter and taking the first 100 words if it is longer. We have realized after the evaluation that this value might be not the best option for short texts, such as tweets, and we plan to optimize this parameter empirically. To avoid overfitting, we first apply dropout (Srivastava et al., 2014), which randomly sets to zero the activation values of $x\%$ of the neurons in a given layer (in this section, $x = 50$).

Convolutional Layer

A convolutional layer exploits local correlations in the input data. In the case of text as input, this translates into extracting correlations between groups of word or character n-grams in a sentence. To do so, each hidden unit of the CNN will only respond (activate) to a specific continuous slice of the input text. This is implemented on `http://keras.io` using convolutional operations with m convolutional filters of width f separately applied to the input, obtaining m representations of this input usually known as feature maps.

Formally, let $T \in \mathbb{R}^{|t| \times n}$ be the matrix embedding for the tweet t and let $F \in \mathbb{R}^{f \times n}$ be a filter, the output of a wide convolution is a matrix $C \in \mathbb{R}^{m \times (|t|+f-1)}$, where each $c_i \in \mathbb{R}^{|t|+f-1}$ is defined as:

1 http://nlp.stanford.edu/data/glove.twitter.27B.zip

$$C_i = \sum_{j,k} T_{[i-f+1:i,:]} \otimes F \tag{17}$$

and where $\otimes$ is the element-wise multiplication, $1 < i < m$; and j and k are the rows and columns of the matrix $T_{[i-f+1:i,:]} \otimes F \in \mathbb{R}^{f \times n}$. The non valid rows of T ($T_{(i,:)}$ with $i < 0$) are set to zero.

Following Severyn and Moschitti (2015), in this paper we chose $f = 5$ and $m = 300$. We also rely on $ReLU(x) = max(0, x)$ as the non-linear activation function. To avoid overfitting we incorporate a L2 regularization of 0.0001. After that, a max pooling layer selects $max(ReLU(c_i))$ for each feature map.

Output layer

The output of the pooling layer is then passed to a fully connected layer ($\mathbb{R}^{100}$). We add again dropout (50%) and a ReLU (LeCun, Bengio, and Hinton, 2015) as the activation function. Finally, an additional fully connected layer reduces the dimensionality of the input to fit the output (number of classes) and as the final step we apply a softmax function (Equation 18) to make the final prediction:

$$softmax = P(y = j|x, b) = \frac{e^{xw+b}}{\sum_{n=0}^{N} e^{xw+b}} \tag{18}$$

where N is the number of classes.

Current limitations

Obtaining an accurate deep neural network can be a very slow process. Hyper-parameter engineering is often needed, but training a single DNN with its hyper-parameters can be painfully slow without enough computational resources. Additionally, *distant supervision* is also recommended to pretrain the network (Go, Bhayani, and Huang, 2009; Severyn and Moschitti, 2015). These two issues act as limitations that we could not overcome at the moment. We did try pretraining, but at the moment, we did not achieve improvements over the CNN without pretraining. A preliminary analysis suggests that: (1) we need more tweets to exploit distant supervision, (2) fine hyper-parameter engineering needs to be explored to ensure that the fine-tuning on the labeled data does not completely overwrite what the network has already learned and (3), it is easy to collect tweets for analysis into 2 classes, but downloading non-noisy tweets for analysis into 3 and 5 classes is a more challenging issue.

In the following section we show how to exploit the hidden activation values of our deep learning model as part of a supervised system (Poria, Cambria, and Gelbukh, 2015), when pretraining and fast hyper-parameter engineering are not feasible options.

11.1.2 *Classifier*

We rely on a support vector machine (SVM), in particular on a liblinear (Fan et al., 2008) implementation with L2-regularization, to train our supervised model.[2] As features, we started testing some of those from (Vilares et al., 2014a), an English implementation of the system described in Chapter 6, using the total occurrence as the weighting factor. Information gain (IG) is used in all cases. Thus, before training our classifier we run an information gain algorithm to remove all irrelevant features, i. e. those where IG=0:

- *Words* (W): Each single word is considered as a feature to feed the supervised classifier.
- *Psychometric properties* (P): Features extracted from psychological properties coming from LIWC (Pennebaker, Francis, and Booth, 2001) that relate terms with psychometric properties (e. g. *anger* or *anxiety*) or topics (e. g. *family* or *religion*).
- *Part-of-speech tags* (T).

Additionally, we have included:

- *The last word of the tweet* (LW): The last term of each tweet is used as a separate feature.
- *The psychometric properties of the last word of the tweet* (LP).
- *Hidden activation values from the* CNN (HV): We take the hidden activation values of the last hidden layer.
- *Features extracted from sentiment dictionaries*: We extract the total, maximum, minimum and last sentiment score of a tweet from the Sentiment140 (Hu and Liu, 2004; Mohammad, Kiritchenko, and Zhu, 2013; Taboada et al., 2011) subjective lexica.

11.1.3 *Dataset*

For all subtasks, three official splits are provided: training, development and development test sets. In this section, we use the training and development sets for training, and the development test set for evaluation.[3]

11.1.4 *Experimental results*

Table 60 shows the experimental results for classification into two classes obtained using the SVM with different feature sets and the CNN. The neural network outperforms most of the SVM approaches. Only

2 We used Weka (Hall et al., 2009) to build the models.

3 For classification into 3 polarities, we include the training set of SemEval 2013 as part of our training set and its development set as a part of our collection for tuning.

when we combine a number of linguistic features with the hidden activation values and we weight the classes, we obtain an improvement over the CNN. We believe that by applying fine hyper-parameter tuning on the CNN we will be able to further improve these results. Similar conclusions can be extracted from the classification into three classes, whose results are shown in Table 61. Finally, Table 62 details the results for the five categories classification subtask. In this particular case, Nakov et al. (2016a) used *macroaveraged mean absolute error* (MAE) (Equation 19) as the official ranking metric. Also, in this latter case the neural network does not perform as well as in previous scenarios.

$$MAE = \frac{1}{|N|}\sum_{j=1}^{|N|}\frac{1}{|T_{c_j}|}\sum_{x_i \in T_{c_j}} |h(x_i) - y_i| \tag{19}$$

where:

- N is the total number of classes.
- y_i is the true label of x_i.
- $h(x_i)$ is the predicted label for x_i.
- T_{c_j} denotes the set of test documents whose true class is c_j.
- $|h(x_i) - y_i|$ is the distance between the predicted and the gold class. The classes are considered as ordered and discrete values (e.g. NEG+(0), NEG(1), NEU(2), POS(3) and POS+(4)).

Features	pos-r	neg-r	Macro avg. r
HV ∪ P ∪ D ∪ LW ∪ LP ∪ FT*	0.721	**0.803**	**0.762**
HV ∪ P ∪ D ∪ LW ∪ LP ∪ FT	0.856	0.581	0.719
HV	0.864	0.560	0.712
P	0.953	0.192	0.573
W	0.969	0.162	0.566
D	0.892	0.249	0.564
CNN	0.802	0.671	0.737

Table 60: *Classification into two classes* using the SemEval 2016 development test set. We include feature models that include hidden activation values (HV), words (W), psychometric (P), sentiment dictionaries (D), last word of the tweet (LW) and last psychometric properties (LP). The dot indicates a model that combines those features. '*' indicates a model where the class weights have been tuned. We compared them against our CNN.

With respect to SVM-specific parameter optimization, cost parameter (C) and class weights (W):

Features	posf1	neu-f1	neg-f1	Macro avg. f1
HV ∪ P ∪ FT ∪ D ∪ LW ∪ LP*	**0.676**	0.520	**0.538**	**0.598**
HV ∪ P ∪ FT ∪ D ∪ LW ∪ LP	0.664	0.565	0.483	0.576
HV	0.659	0.574	0.469	0.564
P	0.620	0.524	0.353	0.487
W	0.611	**0.614**	0.327	0.469
D	0.613	0.553	0.302	0.458
CNN	0.674	0.493	0.489	0.582

Table 61: *Classification into three classes* using both the SemEval 2016 development test set and the SemEval 2013 development set3. Macro-averaged F1 of positive and negative tweets is used to rank the models.

Features	pos+-f1	pos-f1	neu-f1	neg-f1	neg+-f1	mae
HV ∪ P ∪ FT ∪ D ∪ LW ∪ LP*	**0.277**	0.621	**0.439**	0.296	**0.237**	**0.83**
HV ∪ P ∪ FT ∪ D ∪ LW ∪ LP	0.098	0.689	**0.439**	**0.304**	0.063	0.93
HV	0.000	**0.690**	0.417	0.277	0.000	0.95
P	0.000	0.676	0.246	0.070	0.000	1.21
W	0.016	0.674	0.227	0.059	0.000	1.28
CNN	0.000	0.703	0.361	0.229	0.000	1.03

Table 62: *Classification into five classes* using the SemEval 2016 development test set. Macro-averaged absolute error (MAE) is used to rank the models. F1-measure is used to show the performance over each class.

- *2 classes*: C=0.005, $w_{negative}$=2.25 and $w_{positive}$=0.25.
- *3 classes*: C=0.0001, $w_{positive}$=0.5, $w_{neutral}$=0.4 and $w_{negative}$=2.
- *5 classes*: C=1, $w_{strong\ negative}$=5.5, $w_{negative}$=1, $w_{strong\ positive}$=1.5, $w_{positive}$=0.25 and $w_{neutral}$=0.5.

11.2 CONCLUSION

This chapter has described our participation at the SemEval 2016. We included in our sentiment analysis pipeline a convolutional neural network, trained it and then used deep features to feed an external classifier, which served to push up the performance of a model based on the features introduced in Chapter 6. In light of the results obtained, we can state that our convolutional network did not perform well as a standalone classifier, due to the lack of enough pretraining data and resources to carry out an exhaustive architecture engineering process, but the activation values of its hidden values seemed to be really useful as deep features for an external classifier.

Part V

CONCLUSION

12

CONCLUSION

The main goal of this dissertation was to provide insights into the area of *sentiment analysis* and *polarity classification* (Pang and Lee, 2008). In particular, we addressed two related challenges: (1) the development of methods to handle semantic compositionality at the phrase and sentence levels, i. e. the ability to accurately compound the sentiment where the global sentiment might be different or even opposite to the one coming from each of their individual components and (2) their application to multilingual environments. We explored a variety of approaches to achieve these goals, including knowledge-based approaches (Chapters 3, 4 and 5) and machine learning models (Chapters 6 and 7).

In Part ii we proposed different approaches to address semantic compositionality by relying on knowledge-based models. We first focused on the Spanish language and improved SentiStrength (Thelwall, Buckley, and Paltoglou, 2012), a purely lexical multilingual system available for Spanish among other languages, which can handle relevant linguistic constructions to determine the sentiment of a text, such as intensification or negation, by relying on window-based rules. It also includes different configuration options to tackle other phenomena that can be relevant for the challenge at hand (e. g. the influence of exclamation and interrogation marks). We created a corpus according to the *sentistrength* criteria, added new lexical resources to improve the coverage of the subjective dictionaries, and evaluated different configuration setups. The experiments showed that the new Spanish version clearly improved over the existent baseline and that the impact of disabling the treatment of individual phenomena was in general small. The main advantage of Spanish SentiStrength is its simplicity and robustness to perform fast large-scale data analysis in real time, consuming few resources. However, this approach lacked the needed capacities to manage linguistic cases whose scope is non-local and it cannot be effectively captured by simple window-based rules, an issue that is common when dealing with natural language.

To overcome this limitation of the traditional lexical approach, we built a syntactic model for monolingual sentiment analysis, using as case of study the Spanish language. We identified a set of syntactic patterns over Spanish dependency trees annotated according to the Ancora guidelines (Taulé, Martí, and Recasens, 2008), that helped us define rules to manage different phenomena and identify their scope, which can be either fairly local (e. g. intensification) or potentially unrestricted (e. g. negation and subordinate adversative clauses).

We evaluated this model over general- and specific-domain corpora. Experiments on the general-domain corpus showed that our model obtained state-of-the-art results. With respect to the specific-domain corpora, it was observed that adapting the semantic orientation of the subjectivity lexica was first required. To do this, we proposed a semi-automatic method to enrich and adapt such semantic dictionaries to a particular domain, obtaining as a result in the target corpora a performance similar to the one obtained with state-of-the-art machine learning tools. Overall, the results reinforced the practical advantages against traditional lexicon-based models, but in contrast to Spanish SentiStrength, it was heavily dependent, not just on the language, but also on the dependency structure annotation criteria, which complicated its adaptation to multilingual environments.

To counteract this, we introduced a theoretical formalism for compositional operations, allowing the creation of arbitrarily complex rules to tackle relevant phenomena for sentiment analysis, for any language and syntactic dependency annotation. As a result, we can handle multilinguality as easily as SentiStrength does. To prove its usefulness, we implemented and evaluated a set of practical universal operations defined under the universal guidelines of Petrov, Das, and McDonald (2011), McDonald et al. (2013) and Nivre et al. (2016). We provided an evaluation on a total of 7 languages split betwwen two different configurations: (1) different monolingual models (English, German and Spanish) that shared the compositional operations and (2) a single multilingual model that can analyze five Iberian languages (Basque, Catalan, Galician, Portuguese and Spanish) where in addition to the compositional operations, the subjectivity lexica and tagging and parsing models are also shared. The experimental results showed that a practical implementation of the formalism outperformed two of the most commonly used unsupervised systems, proved the universality of the model's compositional operations across different languages and reinforced the usefulness of our approach on domain-transfer applications in comparison to supervised models.

In Part iii, we have taken a different perspective, relying on machine-learning models and focusing on tweets. We first focused on Spanish tweets and proposed different sets of features that related lexical, syntactic, psychological and semantic information to then measure how these affected polarity classification on tweets. One of the novelties in this respect came from the use of enriched generalized dependency triplets, a representation of syntactic relations that connect pairs of words in dependency trees, where the head and the dependent terms can be represented by high-level abstract concepts. This dissertation also illustrated how large a corpus should be in order to take advantage of these syntactic features and also how the size of the training collection influences traditional sets of features. To the best of our

knowledge, this was the first wide evaluation of the effectiveness of using these features, both in isolation and in combination, on a corpus of Spanish Twitter messages.

In a similar line to what we did in the case of knowledge-based approaches, we then explored how such machine learning approaches could be adapted to work on multilingual environments. In particular, we used as experimental framework English and Spanish tweets and their occurrence on monolingual, multilingual and code-switching scenarios. To evaluate the code-switching scenario, we have presented the first code-switching Twitter corpus for multilingual sentiment analysis, composed of tweets that merge English and Spanish terms. We evaluated and compared three different perspectives to perform multilingual polarity classification under these environments: (a) a multilingual model trained on a corpus that fuses two monolingual corpora, according to level 2 (situation refinement) of application of information fusion techniques to the sentiment analysis pipeline, described by Balazs and Velásquez (2016), (b) a dual monolingual model and (c) a simple pipeline which used language identification techniques to determine the language of unseen texts. The experimental results showed that the purely multilingual approach (a) performed very robustly under monolingual, multilingual and code-switching corpora, concluding that is possible to teach a supervised model at least one additional language without significant loss of performance.

In Part iv we described our results at other challenges related to data analysis and evaluation campaigns. We first showed how the model described in Chapter 6 could perform very competitively in the area of multi-label topic classification. We studied which features were more relevant for the purpose at hand and evaluated them using standard metrics for multi-label classification problems. We then described how we addressed the problem of real-time political analysis in the context of Spain, by relying on Spanish SentiStrength (Chapter 3). We crawled Twitter seeking tweets containing mentions to the most popular politicians in Spain, analyzed them with SentiStrength and showed how the positive and negative scores given by the system were in line with the levels of popularity provided by official and unofficial polls. However, we also illustrated how such scores seem to be helpful in order to make electoral predictions.

We then reviewed our participation in evaluation campaigns. We described our participation at RepLab 2014, a competitive evaluation for reputation classification and author ranking (in terms of influential and non influential authors). Again, we showed how our machine learning approach could perform robustly in these scenarios too. Last but not least, we detailed the deep learning approach used at SemEval 2016. We trained a convolutional neural network to then extract the activation values of the hidden layers and used them as deep

features for an external classifier, showing how such features could significantly increase the performance of the core machine-learning model used throughout this dissertation.

12.1 FUTURE WORK

This dissertation has focused on sentiment analysis and polarity classification, with a special emphasis on handling semantic compositionality and multilinguality. In the near future, we plan to move one step forward in such areas:

- In very recent years, deep learning has impacted many areas related to artificial intelligence, and sentiment analysis has not been a stranger to this phenomenon. This dissertation already presented experimental results on standard evaluation campaigns using neural networks, but such results were still preliminary. In particular, we are interested in determining if the approach described in Chapter 7 could be useful in the context of deep learning classifiers.
- Also, we would like to explore how we could adapt convolutional neural networks to individually extract the part of texts containing the most subjective phrase in English and Spanish and then make different predictions for each language. This could be helpful for sociolinguistics, showing how code-switching speakers use subjective language.
- In the intersection of multilingual sentiment analysis and neural network architectures relying on distributed inputs, the field of multilingual embeddings plays an important role. Mapping multilingual embeddings to the same multidimensional space might have potential advantages. For example, we could train a neural network using English embeddings (for which we have a large amount of labeled data) and then use the trained network to make predictions on a new language (for which we have trained embeddings in the same multidimensional space, but not much labeled data).
- Aspect-based sentiment analysis has arisen the interest of the NLP community recently (Pontiki et al., 2014). We would like to explore if the formal approach described in Chapter 5 could be successfully transfered to carry out aspect-based sentiment analysis. Also, recent models in this area can help to inspire us in this subfield of sentiment analysis (Wang et al., 2016).

Last, the results presented in this dissertation can help other computer science areas, such as computational social science, information retrieval and recommender systems.

Appendices

A

POLLED POLITICIANS SAMPLED IN CHAPTER 9

The appendix describes the parties and politicians studied in Chapter 9 and their approximate number of followers and responsabilities at beginning of 2014:

- Partido Popular (@PPopular) 189,000 followers
  - –Mariano Rajoy (@marianorajoy): The Prime Minister with 650 000 followers.
  - –Soraya Sáenz de Santamaría (@Sorayapp): The Deputy Prime Minister with 154 000 followers.
  - –María Dolores de Cospedal (@mdcpospedal): The PP Secretary-General with 88 900 followers.
  - –Esperanza Aguirre (@EsperanzAguirre): President of the PP Madrid federation with 245 000 followers.
  - –Cristina Cifuentes (@ccifuentes): Delegate of the Spanish government in Madrid with 65 500 followers.
- Partido Socialista Obrero Español (@PSOE) 195 000 followers
  - –Pedro Sánchez (@sanchezcastejon): The leader and Secretary-General of PSOE with 112 000 followers.
  - –César Luena (@cesarluena): The PSOE Secretary and deputy leader with 9 848 followers.
  - –Susana Díaz (@_susanadiaz): President of Andalucia with 44 400 followers.
  - –Carme Chacón (@carmechacon): Former Minister of Defence with more than 87 900 followers.
  - –Elena Valenciano (@ElenaValenciano): Head of the PSOE 2014 European election list with 21 300 followers.
- Podemos (@ahorapodemos) 482,000 followers
  - –Pablo Iglesias (@_Pablo_Iglesias): Leader and Secretary-General of the party with 739 000 followers.
  - –Juan Carlos Monedero (@MonederoJC): Program Secretary of Podemos with 128 000 followers.
  - –Iñigo Errejón (@ierrejon): Secretary of Politics with 145 000 followers.
  - –Pablo Echenique (@pnique): Representative of Podemos in the European parliament with 95 800 followers.

  - –Teresa Rodríguez (@TeresaRodr_): A European parliamentary member with 64 700 followers.
- Izquierda Unida (@unida) 124,000 followers
  - –Cayo Lara (@cayo_lara): Coordinator of Izquierda Unida and member of the Spanish parliament with 170 000 followers.
  - –Alberto Garzón (@agarzon): Member of the Spanish parliament with 282 000 followers.
  - –Tania Sánchez (@Ainhat): IU candidate for Mayor of Madrid with 84 700 followers.
  - –Gaspar Llamazares (@GLlamazares): Former head of the party with 227 000 followers.
  - –Hugo Martínez Abarca (@hugomabarca): Member of IU with a high Twitter profile and 26 200 followers.
- Unión, Progreso y Democracia (@UpyD) 106 000 followers
  - –Rosa Díez: Party leader without a Twitter account during the polling period.
  - –Toni Cantó (@ToniCanto1): Spanish actor and member of the Spanish parliament with 169 000 followers.
  - –Irene Lozano (@lozanoirene): Member of the Spanish parliament with 16 700 followers.
  - –Carlos Martínez Gorriarán (@cmgorriaran): Member of the Spanish parliament with 21 300 followers.
  - –Beatriz Becerra: European parliamentary member with 6 700 followers.
- Ciudadanos (@CiudadanosCs): 73 800 followers
  - –Albert Rivera (@Albert_Rivera): Founder and president with 141 000 followers.
  - –Luis Salvador (@luissalvador): Candidate for Mayor of Granada and a high Twitter profile with 63 000 followers.
  - –Fran Hervias (@FranHervias): Secretary with 5 000 followers.
  - –Inés Arrimadas (@InesArrimadas): Catalan member of parliament with 7 000 followers.
  - –Javier Nart (@JavierNart): European parliamentary with 18 400 followers.

B

STATISTICS OF THE TOPIC CLASSIFICATION TRAINING SET USED IN CHAPTER 8

Table 63 shows the frequency statistics of the training set of the topic classification corpus used in Chapter 8.

Categories	**%tweets**	**#tweets**
{films}	1.5	107
{films, economy}	0.0	1
{films, entertainment}	0.3	21
{films, entertainment, music}	0.0	1
{films, entertainment, other}	0.0	2
{films, entertainment, politics}	0.0	3
{films, soccer}	0.0	1
{films, music}	0.1	7
{films, other}	1.3	97
{films, other, politics}	0.0	1
{films, technology}	0.1	4
{sports}	1.0	75
{sports, economy}	0.0	2
{sports, entertainment}	0.2	11
{sports, entertainment, music}	0.0	1
{sports, entertainment, other}	0.0	1
{sports, entertainment, other, politics}	0.0	1
{sports, entertainment, politics}	0.0	1
{sports, soccer}	0.1	5
{sports, literature}	0.0	1
{sports, music}	0.1	4
{sports, music, other}	0.0	1
{sports, other}	0.1	8
{sports, politics}	0.0	1
{sports, technology}	0.0	1
{economy}	3.7	267
{economy, entertainment}	0.4	32
{economy, entertainment, other}	0.1	5

Continued on next page

Categories	%tweets	#tweets
{economy, entertainment, other, politics}	0.0	2
{economy, entertainment, politics}	0.5	36
{economy, entertainment, politics, technology}	0.0	1
{economy, entertainment, technology}	0.0	2
{economy, soccer}	0.0	2
{economy, literature}	0.0	1
{economy, literature, politics}	0.0	1
{economy, literature, politics, technology}	0.0	1
{economy, music}	0.0	1
{economy, music, politics}	0.0	1
{economy, other}	0.3	23
{economy, other, politics}	0.4	28
{economy, other, technology}	0.0	1
{economy, politics}	7.3	529
{economy, politics, technology}	0.0	1
{economy, technology}	0.1	5
{entertainment}	11.5	827
{entertainment, soccer}	0.4	30
{entertainment, soccer, music, other}	0.0	1
{entertainment, soccer, other}	0.0	2
{entertainment, literature}	0.3	19
{entertainment, literature, politics}	0.0	1
{entertainment, literature, technology}	0.0	2
{entertainment, music}	0.5	39
{entertainment, music, other}	0.1	5
{entertainment, music, politics}	0.0	1
{entertainment, music, technology}	0.0	1
{entertainment, other}	4.5	328
{entertainment, other, politics}	0.2	13
{entertainment, other, technology}	0.1	4
{entertainment, politics}	3.3	241
{entertainment, politics, technology}	0.1	4
{entertainment, technology}	0.6	40
{soccer}	2.3	166
{soccer, literature}	0.0	1
{soccer, music}	0.1	8

Continued on next page

Categories	%tweets	#tweets
{soccer, music, other}	0.0	1
{soccer, other}	0.4	27
{soccer, politics}	0.1	7
{soccer, technology}	0.0	1
{literature}	0.6	45
{literature, music}	0.0	2
{literature, other}	0.2	14
{literature, politics}	0.2	13
{literature, technology}	0.0	2
{music}	2.8	200
{music, other}	3.9	279
{music, other, politics}	0.0	1
{music, other, technology}	0.0	1
{music, politics}	0.1	5
{music, technology}	0.1	6
{other}	17.3	1 248
{other, politics}	3.0	215
{other, politics, technology}	0.0	1
{other, technology}	0.4	27
{politics}	27.5	1 982
{politics, technology}	0.4	29
{technology}	1.1	83

Table 63: Statistics of the training set used for the topic classification tasks (Chapter 8)

C

STATISTICS OF THE TOPIC CLASSIFICATION TEST SET USED IN CHAPTER 8

Table 64 shows the frequency statistics of the test set of the topic classification corpus used in Chapter 8.

Categories	**%tweets**	**#tweets**
{films}	0.3	203
{films, entertainment}	0.0	13
{films, entertainment, other}	0.0	1
{films, music}	0.0	5
{films, other}	0.6	368
{films, politics}	0.0	5
{films, technology}	0.0	1
{sports}	0.2	106
{sports, entertainment}	0.0	3
{sports, soccer}	0.0	1
{sports, music}	0.0	1
{sports, other}	0.0	20
{sports, politics}	0.0	4
{economy}	2.0	1 209
{economy, entertainment}	0.0	4
{economy, entertainment, politics}	0.0	1
{economy, soccer}	0.0	1
{economy, other}	0.3	195
{economy, other, politics}	0.0	1
{economy, politics}	1.9	1 138
{entertainment}	5.7	3 494
{entertainment, soccer}	0.0	6
{entertainment, literature}	0.0	9
{entertainment, music}	0.0	6
{entertainment, music, other}	0.0	1
{entertainment, other}	2.4	1 486
{entertainment, other, politics}	0.0	3
{entertainment, other, technology}	0.0	3

Continued on next page

Categories	%tweets	#tweets
{entertainment, politics}	0.6	371
{entertainment, technology}	0.0	20
{soccer}	1.2	700
{soccer, music}	0.0	2
{soccer, other}	0.2	95
{soccer, politics}	0.0	17
{soccer, technology}	0.0	1
{literature}	0.1	76
{literature, other}	0.0	7
{literature, politics}	0.0	1
{music}	0.9	545
{music, other}	1.5	924
{music, politics}	0.0	13
{music, technology}	0.0	1
{other}	34.5	20 979
{other, politics}	6.7	4 081
{other, technology}	0.0	27
{politics}	40.2	24 416
{politics, technology}	0.0	16
{technology}	0.4	218

Table 64: Statistics of the training set used for the topic classification tasks (Chapter 8)

D

LONG SUMMARY IN SPANISH / RESUMEN LARGO EN CASTELLANO

Esta tesis presenta nuevas aproximaciones en el ámbito del *análisis del sentimiento* y la *clasificación de polaridad* (Pang y Lee, 2008), consistente en determinar si el sentimiento de una frase, oración o documento refleja una opinión positiva, negativa o neutral.

En la primera parte de este trabajo, se realiza una introducción al área del análisis del sentimiento y se presentan técnicas de procesamiento del lenguaje natural que se usarán en los siguientes capítulos.

En la segunda parte, se presentan métodos basados en conocimiento para calcular la orientación semántica a nivel de oración, que pueden manejar construcciones lingüísticas relevantes en el ámbito que nos ocupa, como la intensificación, la negación o las oraciones subordinadas adversativas.

En tercer lugar, se describe cómo construir clasificadores de polaridad mediante aprendizaje automático utilizando como características de entrada información léxica, sintáctica y semántica.

Por último, se presentan resultados experimentales obtenidos durante el desarrollo de esta tesis en distintas competiciones de evaluación internacionales y otras áreas de investigación relacionadas con el tema que nos ocupa, como son el análisis político o de reputación.

D.1 MOTIVACIÓN

Analizar y comprender el contenido subjetivo compartido en las redes sociales por usuarios de diferentes países, culturas y edades se ha convertido en un punto clave para monitorizar la opinión pública sobre toda una variedad de productos, eventos o celebridades. Por ejemplo, a partir de foros de opiniones sobre películas, como FilmAffinity, es posible conocer lo que los espectadores piensan sobre distintos aspectos de una película y tomar una decisión sobre qué ver basándose en sus preferencias personales. A partir de foros relacionados con el ámbito turístico, como TripAdvisor, es posible encontrar toda una variedad de opiniones sobre el alojamiento donde un usuario está planeando pasar sus próximas vacaciones. A partir de redes sociales modernas como Facebook, Twitter o Instagram, podemos conocer la opinión de sus usuarios respecto a una noticia, tendencia o incluso determinar el sentimiento que unas determinadas imágenes o vídeos pueden despertar en la sociedad. Todo esta información puede ser procesada por personas de forma natural para transformarla en conocimiento que pueda responder a preguntas como: *'¿Qué opi-*

na la gente sobre la actuación de Edward Norton en American History X?', *'Me preocupa la comodidad de la cama y la limpieza, ¿debería reservar esta habitación?'*, *'¿Cómo está evolucionando la opinión de la población acerca de Samsung Electronics después del incidente con el Samsung Galaxy Note 7 ?'*.

Antes de la aparición de la Web 2.0, una solución habitual para obtener respuestas a preguntas como estas consistía en delegar en estudios y encuestas. Sin embargo, estas estrategias son normalmente caras, con un alcance limitado y típicamente solo válidas durante un corto período de tiempo. En este contexto, las redes sociales pueden ser una forma efectiva de tener información sobre usuarios (Wang y col., 2012) y de planear estrategias de negocio y mercado (Bae y Lee, 2012; Li y Li, 2013). Sin embargo, monitorizar redes sociales manualmente presenta numerosos obstáculos. Por un lado, la gran cantidad de opiniones expresadas hace que la observación y análisis manual sea inviable. Además, estudios previos han demostrado que delegar este tipo de análisis en intuiciones humanas en vez de análisis automáticos puede resultar en la extracción de indicadores de sentimiento sesgados por experiencias personales (Pang, Lee y Vaithyanathan, 2002).

En este contexto, el *análisis del sentimiento* es un campo de investigación enmarcado en el análisis automático del contenido subjetivo, donde una de las subtareas que ha logrado una mayor popularidad consiste en clasificar el *sentimiento* o la *polaridad* de un texto como positivo o negativo, aunque es común incluir categorías adicionales para distinguir también textos objetivos o para diferenciar la fuerza de las opiniones.

Esta tesis se centra en obtener el sentimiento de una frase, oración o documento desde un enfoque basado en procesamiento del lenguaje natural. En concreto, se hace especial énfasis en métodos capaces de manejar la *semántica composicional*, es decir, métodos con la habilidad de componer el sentimiento de oraciones donde el sentimiento global puede ser diferente, o incluso opuesto, del que se obtendría para cada uno de sus términos de manera individual. Por ejemplo, en la oración *'Él no es muy guapo, pero tiene algo que realmente me gusta'*, queremos diseñar algoritmos con la habilidad de inferir que la palabra *'muy'* enfatiza la palabra *'guapo'*, *'no'* afecta a la expresión *'muy guapo'*, y *'pero'* decrementa la relevancia de *'Él no es muy guapo'* e incrementa la de *'tiene algo que realmente me gusta'*. También se presta especial atención en cómo los métodos aquí propuestos pueden ser adaptados a entornos multilingües.

D.2 CONTENIDO DE LA TESIS

En esta sección se hace un resumen en castellano del contenido de cada uno de los capítulos de esta tesis.

PARTE I

En la Parte i además de la introducción se presentan técnicas de procesamiento del lenguaje natural que se usan a lo largo de la tesis.

CAPÍTULO 2

En particular, en el Capítulo 2 se presta especial atención en cómo construir analizadores sintácticos y etiquetadores morfológicos que puedan analizar múltiples idiomas sin necesidad de aplicar ninguna técnica previa de identificación del idioma, algo muy útil en el ámbito del análisis del sentimiento en entornos multilingües, incluyendo escenarios de *code-switching*. En este contexto, en los preliminares se introduce una aproximación para entrenar analizadores sintácticos lexicalizados usando corpus bilingües donde se juntan varios *treebanks* monolingües, que tienen sus anotaciones armonizadas. Como resultado se obtienen analizadores sintácticos que pueden analizar oraciones en cualquiera de los idiomas para los que fueron entrenados, o incluso oraciones que mezclan varios de ellos. Los resultados de esta parte de la tesis prueban que estos analizadores sintácticos pueden ser realmente competitivos, y para varios pares de idiomas, el analizador bilingüe no solo preserva el rendimiento sino que incluso se alcanza una mejora significativa sobre el correspondiente analizador monolingüe.

PARTE II

En la Parte ii se presentan aproximaciones para manejar la semántica composicional a partir de modelos basados en conocimiento.

CAPÍTULO 3

En primer lugar, el Capítulo 3 se centra en el español y el sistema SentiStrength (Thelwall, Buckley y Paltoglou, 2012), un software multilingüe basado en lexicones que puede manejar construcciones lingüísticas relevantes a la hora de obtener el sentimiento de un texto, como la intensificación y la negación, mediante reglas basadas en un alcance puramente léxico, sin tener en cuenta la estructura sintáctica de la oración. Este sistema también incluye diferentes opciones de configuración para tratar otros fenómenos comunes en textos web que pueden influir en el análisis de polaridad (por ejemplo, la in-

fluencia de los signos de exclamación e interrogación). Como rasgo particular, SentiStrength produce una doble salida, que a lo largo de este trabajo se denota con la notación *sentistrength*, donde cada texto recibe dos puntuaciones: una positiva y otra negativa. Para mejorar dicho sistema para el español, primero se ha creado un corpus de acuerdo a dicho estilo de anotación, se han incorporado más recursos léxicos para mejorar la cobertura de los diccionarios subjetivos, y se han evaluado distintas configuraciones que permiten deshabilitar el tratamiento de distintos fenómenos web. Los experimentos muestran que la nueva versión de SentiStrength para el castellano mejora con claridad el modelo existente y que el impacto de deshabilitar opciones de configuración para ignorar ciertos fenómenos comunes en textos web es, en general, pequeño. La principal ventaja de SentiStrength es su sencillez y robustez para llevar a cabo análisis de grandes cantidades de datos en tiempo real, consumiendo pocos recursos; así como su fácil adaptación a distintos idiomas, que únicamente requiere de la creación de nuevos recursos léxicos para el idioma deseado. Sin embargo, este modelo también presenta algunos inconvenientes. En concreto, no dispone de las capacidades necesarias para manejar fenómenos lingüísticos cuyo alcance no es local y no puede ser identificado de forma precisa usando solo reglas léxicas, algo muy habitual al trabajar con lenguajes naturales.

CAPÍTULO 4

Para superar esta limitación de los sistemas puramente léxicos, en el Capítulo 4 construimos un modelo sintáctico para análisis del sentimiento monolingüe, usando como caso de estudio el español. En primer lugar, identificamos un conjunto de patrones sintácticos presentes en árboles de dependencias para este idioma, anotados conforme a las normas de Ancora (Taulé, Martí y Recasens, 2008), que nos permiten definir reglas para detectar fenómenos lingüísticos relevantes y su alcance, que puede ser local (por ejemplo, la intensificación) o lejano (por ejemplo, la negación y oraciones subordinadas adversativas). Este modelo se evalúa tanto en colecciones de carácter general como de dominio específico. En el primer caso, los experimentos muestran que nuestro modelo obtiene resultados superiores a los de trabajos previos en el mismo corpus. Se ilustra cómo hay determinadas temáticas que son más sencillas (por ejemplo, el turismo), dado que las orientaciones semánticas de los términos subjetivos habituales se corresponden en general con la percepción genérica que la población tiene sobre esa palabra; mientras que otras son notablemente más complicadas de tratar con este tipo de métodos (por ejemplo, críticas sobre películas), dado que la percepción de muchos de estos términos tiene una orientación semántica opuesta a la habitual. Con respecto a la evaluación sobre dominios específicos, se hace

manifiesta la necesidad de adaptar primero la orientación semántica de los términos de los diccionarios subjetivos usados por nuestro sistema. Para ello, se propone un método semi-automático de adaptación al dominio, ilustrando cómo de esta manera es posible obtener resultados similares a los obtenidos por un sistema de aprendizaje automático. En términos generales, los resultados demuestran que Íla aproximación sintáctica propuesta presenta ventajas con respecto a los modelos léxicos, pero también algunos inconvenientes. Entre ellos cabe destacar por ejemplo que, a diferencia de SentiStrength, el modelo propuesto es dependiente, no solo del idioma, sino también del criterio usado para la anotación de los árboles de dependencias, lo que complica su adaptación a entornos multilingües.

CAPÍTULO 5

El Capítulo 5 lidia con esta debilidad de los sistemas sintácticos e introduce un formalismo teórico para la definición de operaciones composicionales, un concepto que permite la creación de reglas arbitrariamente complejas para tratar construcciones lingüísticas relevantes en el ámbito del análisis del sentimiento, para cualquier idioma y estilo de anotación sintáctica, de manera que sea posible trabajar en entornos multilingües con la misma sencillez que SentiStrength permite. Para probar la utilidad del formalismo propuesto, implementamos y evaluamos un conjunto práctico de operaciones coherentes con los estilos de anotación de Petrov, Das y McDonald (2011), McDonald y col. (2013) y Nivre y col. (2016), con el objetivo de tratar la intensificación, negación, oraciones subordinadas adversativas y el *irrealis* (en particular el condicional *'si'*). A continuación se lleva a cabo una evaluación sobre un total de 7 idiomas entre dos configuraciones distintas: (1) diferentes modelos monolingües (para inglés, alemán y español) que comparten el mismo conjunto de operaciones composicionales y (2) un único modelo multilingüe que puede analizar cinco idiomas oficiales en la península ibérica (catalán, español, gallego, portugués y vasco), donde además de las operaciones composicionales también se comparten los diccionarios subjetivos así como el etiquetador morfológico y el analizador sintáctico. Los resultados experimentales muestran: (1) cómo un modelo que implementa dichas operaciones puede obtener mejor rendimiento que los sistemas léxicos más utilizados, (2) que las operaciones composicionales pueden ser compartidas entre varios idiomas y (3) el buen rendimiento de nuestro modelo en entornos dispares, en comparación con modelos supervisados.

PARTE III

En la Parte iii, se aborda el problema del análisis del sentimiento desde un enfoque basado en aprendizaje automático y centrándonos en la clasificación de tuits.

CAPÍTULO 6

En primer lugar y de manera análoga al Capítulo 3, el Capítulo 6 se centra en mensaje escritos en español, en particular tuits, y propone diferentes conjuntos de características que contienen información léxica, sintáctica y semántica, observando cómo su uso influye negativa o positivamente en la clasificación de la polaridad. Se presentan también las *tripletas sintácticas generalizadas enriquecidas*, una representación sintáctica que conecta pares de palabras extrayendo su relación en el árbol de dependencias, donde el padre o el término dependiente pueden ser representados por conceptos más generales y más abstractos. También mostramos cómo de grande debe ser el conjunto de entrenamiento para que este tipo de características tengan un impacto positivo en el rendimiento y no se vean afectadas por problemas de dispersión.

CAPÍTULO 7

De manera similar a lo realizado con las aproximaciones basadas en conocimiento, el Capítulo 7 explora cómo modelos de aprendizaje automático pueden ser adaptados para funcionar de manera efectiva en entornos multilingües. En particular, consideramos escenarios con mensajes en español e inglés, incluyendo textos de *code-switching* (término que se utiliza para denotar textos que contienen palabras en dos o más idiomas). Para este último escenario, anotamos un corpus de tuits que mezclan términos en ambas lenguas. A continuación, evaluamos y comparamos tres perspectivas distintas para llevar a cabo la clasificación de polaridad multilingüe: (a) un modelo multilingüe entrenado en un corpus que fusiona dos colecciones monolingües, (b) un doble modelo monolingüe, (c) una aproximación que usa técnicas de identificación del idioma para determinar en primer lugar el idioma del texto en cuestión y saber a continuación qué clasificador monolingüe debe usarse. Los resultados experimentales muestran que el modelo multilingüe (a) es capaz de obtener resultados robustos en todas las configuraciones de evaluación, lo que nos permite concluir que es posible entrenar un modelo supervisado añadiendo al menos un idioma adicional, sin que ello suponga una pérdida significativa de rendimiento.

PARTE IV

En la Parte iv, se describe nuestra participación en competiciones de evaluación internacionales y otras tareas relacionadas con el análisis automático de textos.

CAPÍTULO 8

El Capítulo 8 muestra cómo el modelo basado en aprendizaje automático presentado en la Parte iii también puede ser utilizado para tareas de clasificación de temáticas, donde un mensaje puede pertenecer a uno o más temas. Twitter es un servicio donde millones de usuarios comparten sus opiniones, como explicamos anteriormente. Sin embargo, la red social no provee una manera eficaz de clasificar que tuits pertenecen a una determinada temática, lo que complica que se puedan utilizar técnicas de análisis del sentimiento sobre Twitter si el objetivo es monitorizar una entidad enmarcada en una temática concreta. En esta tesis, este problema se aborda desde una perspectiva de procesamiento del lenguaje natural y aprendizaje supervisado. Se estudia qué características son las más beneficiosas para esa tarea, evaluando su rendimiento respecto a distintas métricas estándar para clasificación multi-etiqueta.

CAPÍTULO 9

A continuación, el Capítulo 9 ilustra cómo el análisis del sentimiento puede ser utilizado para medir la reputación que los usuarios tienen de políticos en Twitter en tiempo real. Para ello, se ha descargado una gran cantidad de tuits que contienen menciones de los políticos más populares en España, y los hemos analizado con SentiStrength. Los resultados muestran que los niveles de positividad y negatividad son coherentes con los niveles de popularidad que se obtienen a través de diferentes encuestas. Sin embargo, dichas puntuaciones parecen no ser útiles si el objetivo es realizar predicciones electorales.

CAPÍTULO 10

El Capítulo 10 describe nuestra participación en la tarea RepLab 2014 (Amigó y col., 2014), una evaluación competitiva centrada en clasificación de reputación y *ranking* de autores (en términos de su nivel de influencia). Determinar cómo la sociedad relaciona a entidades con conceptos como rendimiento, innovación o liderazgo está estrechamente relacionado con la percepción que ésta tiene sobre ellas, y por consiguiente con el análisis del sentimiento. Este capítulo explora de nuevo cómo utilizar modelos supervisados y características de entrada introducidas en capítulos previos, para la tarea en cuestión, además de dos aproximaciones simples pero efectivas para hacer un *ranking* de los autores. El modelo creado obtuvo el tercer puesto en la

tarea de clasificación de reputación y un segundo puesto en la tarea de *ranking* de autores.

CAPÍTULO 11

Por último, en el Capítulo 11 se describe un modelo basado en redes neuronales que fue utilizado en la competición SemEval 2016 (Nakov y col., 2016b). En particular, primero entrenamos una red convolucional que luego fue utilizada para extraer los valores de activación de sus capas ocultas para usarlos como características de entrada a un clasificador supervisado como los utilizados en capítulos previos, mostrando cómo este tipo de características pueden aumentar significativamente el rendimiento del modelo supervisado. El modelo obtuvo el segundo y cuarto puesto en las tareas de clasificación en dos y cinco polaridades, respectivamente.

D.3 CONTRIBUCIONES

El trabajo realizado en esta tesis ha contribuido al avance del análisis del sentimiento y otras tareas relacionadas con el análisis de textos, mediante la definición formal de técnicas de procesamiento del lenguaje natural, su implementación como parte de bibliotecas, y la construcción de recursos lingüísticos. En particular, las principales contribuciones de esta tesis han sido:

- Un conjunto pre-entrenado de analizadores sintácticos bilingües y multilingües que pueden ser usados para llevar a cabo análisis de sentimiento en un entorno multilingüe.
- Una versión de SentiStrength para el español, un sistema léxico con capacidades multilingües especialmente utilizado en textos cortos (como los compartidos en plataformas como Youtube o Twitter), así como un corpus de tuits en español anotados siguiendo el estilo *sentistrength*.
- Un sistema sintáctico para el análisis en castellano. El sistema funciona para textos escritos en español que han sido previamente procesados por un sistema que obtiene su árbol sintáctico según la estructura definida por Ancora.
- Un formalismo para definir operaciones composicionales, permitiendo la creación de reglas arbitrariamente complejas para manejar fenómenos relevantes en el ámbito del análisis del sentimiento, para cualquier idioma o estilo de anotación de árboles sintácticos. Llevamos a cabo una implementación de dicho formalismo y evaluamos un conjunto de operaciones composicionales que pueden ser compartidas entre varios idiomas.
- Nuevos métodos para clasificar la polaridad de tuits escritos usando aprendizaje automático y características de entrada que represen-

tan distintos niveles de información lingüística (léxica, sintáctica y semántica), en entornos monolingües, multilingües y de *code-switching*.

- El primer corpus de *code-switching* con textos que mezclan inglés y español, anotado siguiendo el estilo de *sentistrength* y también siguiendo una escala trinaria donde se considera que un texto puede ser positivo, negativo o neutro.
- Un análisis en tiempo real de los principales políticos españoles y lo que los usuarios de Twitter dicen de ellos, en términos de positividad y negatividad, comparando los resultados con los de diferentes encuestas.

BIBLIOGRAPHY

Agarwal, A, B Xie, I Vovsha, O Rambow, and R Passonneau (2011). "Sentiment analysis of Twitter data." In: *Proceedings of the Workshop on Languages in Social Media*. LSM '11. Stroudsburg, PA, USA: ACL, pp. 30–38. ISBN: 978-1-932432-96-1.

Agerri, R., J. Bermudez, and G. Rigau (2014). "IXA pipeline: Efficient and Ready to Use Multilingual NLP tools." In: *Proceedings of the 9th edition of the Language Resources and Evaluation Conference (LREC 2014)*, pp. 3823–3828.

Aldayel, Haifa K and Aqil M Azmi (2016). "Arabic tweets sentiment analysis - a hybrid scheme." In: *Journal of Information Science* 42.6, pp. 782–797.

Alonso, Miguel A., Carlos Gómez-Rodríguez, David Vilares, Yerai Doval, and Jesús Vilares (2015). "Seguimiento y análisis automático de contenidos en redes sociales." In: *Actas: III Congreso Nacional de i+d en Defensa y Seguridad, DESEi+d 2015*. Marín, Spain: Centro Universitario de la Defensa de Marín, pp. 899–906.

Amigó, Enrique, Jorge Carrillo de Albornoz, Irina Chugur, Adolfo Corujo, Julio Gonzalo, Tamara Martín-Wanton, Edgar Meij, Maarten de Rijke, and Damiano Spina (2013). "Overview of RepLab 2013: Evaluating Online Reputation Monitoring Systems." In: *Information Access Evaluation. Multilinguality, Multimodality, and Visualization - 4th International Conference of the CLEF Initiative, CLEF 2013, Valencia, Spain, September 23-26, 2013. Proceedings*. Ed. by Pamela Forner, Henning Müller, Roberto Paredes, Paolo Rosso, and Benno Stein. Vol. 8138. Lecture Notes in Computer Science. Berlin and Heidelberg: Springer, pp. 333–352.

Amigó, Enrique, Jorge Carrillo-de Albornoz, Irina Chugur, Adolfo Corujo, Julio Gonzalo, Edgar Meij, Maarten de Rijke, and Damiano Spina (2014). "Overview of Replab 2014: author profiling and reputation dimensions for online reputation management." In: *International Conference of the Cross-Language Evaluation Forum for European Languages*. Springer, pp. 307–322.

Ammar, Waleed, George Mulcaire, Miguel Ballesteros, Chris Dyer, and Noah A. Smith (2016). "Many Languages, One Parser." In: *Transactions of the Association for Computational Linguistics* 4, pp. 431–444.

Andor, Daniel, Chris Alberti, David Weiss, Aliaksei Severyn, Alessandro Presta, Kuzman Ganchev, Slav Petrov, and Michael Collins (2016). "Globally Normalized Transition-Based Neural Networks." In: *Proceedings of the 54th Annual Meeting of the Association for Com-*

putational Linguistics (Volume 1: Long Papers). Berlin, Germany: Association for Computational Linguistics, pp. 2442–2452.

Arakawa, Yui, Akihiro Kameda, Akiko Aizawa, and Takafumi Suzuki (2014). "Adding Twitter-specific features to stylistic features for classifying tweets by user type and number of retweets." In: *Journal of the Association for Information Science and Technology* 65.7, pp. 1416–1423.

Aue, Anthony and Michael Gamon (2005). "Customizing sentiment classifiers to new domains: A case study." In: *Proceedings of recent advances in natural language processing (RANLP)*. Vol. 1. 3.1, pp. 1–2.

Bae, Younggue and Hongchul Lee (2012). "Sentiment analysis of Twitter audiences: Measuring the positive or negative influence of popular twitterers." In: *Journal of the American Society for Information Science and Technology* 63.12, pp. 2521–2535.

Bakliwal, A, P Arora, S Madhappan, N Kapre, M Singh, and V Varma (2012). "Mining Sentiments from Tweets." In: *Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis*. Jeju, Korea: ACL, pp. 11–18.

Balage Filho, P. P., T. AS Pardo, and S. M. Aluísio (2013). "An evaluation of the Brazilian Portuguese LIWC dictionary for sentiment analysis." In: *Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology (STIL)*, pp. 215–219.

Balahur, Alexandra, Zornitsa Kozareva, and Andrés Montoyo (2009). "Determining the polarity and source of opinions expressed in political debates." In: *International Conference on Intelligent Text Processing and Computational Linguistics*. Springer, pp. 468–480.

Balahur, Alexandra and Marco Turchi (2012a). "Comparative Experiments for Multilingual Sentiment Analysis Using Machine Translation." In: *SDAD 2012, The 1st International Workshop on Sentiment Discovery from Affective Data*. Ed. by Mohamed Medhat Gaber, Mihaela Cocea, Stephan Weibelzahl, Ernestina Menasalvas, and Cyril Labbe. Bristol, UK, pp. 75–86.

— (2012b). "Multilingual Sentiment Analysis using Machine Translation?" In: *WASSA 2012, 3rdWorkshop on Computational Approaches to Subjectivity and Sentiment Analysis, Proceedings of the Workshop*. Jeju, Republic of Korea, pp. 52–60.

— (2014). "Comparative experiments using supervised learning and machine translation for multilingual sentiment analysis." In: *Computer Speech and Language* 28.1, pp. 56–75.

Balahur, Alexandra, Marco Turchi, Ralf Steinberger, Jose Manuel Perea-Ortega, Guillaume Jacquet, Dilek Kucuk, Vanni Zavarella, and Adil El Ghali (2014). "Resource Creation and Evaluation for Multilingual Sentiment Analysis in Social Media Texts." In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Ed. by Nicoletta Calzolari (Conference Chair),

Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis. Reykjavik, Iceland: European Language Resources Association (ELRA). ISBN: 978-2-9517408-8-4.

Balazs, Jorge A and Juan D Velásquez (2016). "Opinion mining and information fusion: a survey." In: *Information Fusion* 27, pp. 95–110.

Ballesteros, M and J Nivre (2012). "MaltOptimizer: an optimization tool for MaltParser." In: *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, pp. 58–62.

Banea, Carmen, Rada Mihalcea, and Janyce Wiebe (2010). "Multilingual Subjectivity: Are More Languages Better?" In: *COLING 2010. 23rd International Conference on Computational Linguistics. Proceedings of the Conference*. Ed. by Chu-Ren Huang and Dan Jurafsky. Vol. 2. Beijing: Tsinghua University Press, pp. 28–36.

Banea, Carmen, Rada Mihalceaa, and Janyce Wiebe (2014). "Sense-level subjectivity in a multilingual setting." In: *Computer Speech & Language* 28.1, pp. 7–19.

Barberá, Pablo (2012). "A New Measure of Party Identification in Twitter. Evidence from Spain." In: *2nd Annual General Conference of the European Political Science Association(EPSA)*. EPSA. Berlin.

Barberá, Pablo and Gonzalo Rivero (2012). "¿Un tweet, un voto? Desigualdad en la discusión política en Twitter." In: *I Congreso Internacional en Comunicación Política y Estrategias de Campaña*.

Beek, Leonoor Van der, Gosse Bouma, Rob Malouf, and Gertjan Van Noord (2002). "The Alpino dependency treebank." In: *Language and Computers* 45.1, pp. 8–22.

Benjamini, Yoav and Yosef Hochberg (1995). "Controlling the false discovery rate: a practical and powerful approach to multiple testing." In: *Journal of the royal statistical society. Series B (Methodological)*, pp. 289–300.

Bentham, Jeremy (1789). *The principles of moral and legislation*.

Bermingham, Adam and Alan F Smeaton (2011). "On using Twitter to monitor political sentiment and predict election results." In: pp. 2–10.

Böhmová, Alena, Jan Hajič, Eva Hajičová, and Barbora Hladká (2003). "The Prague dependency treebank." In: *Treebanks*. Springer, pp. 103–127.

Boiy, Erik and Marie-Francine Moens (2009). "A machine learning approach to sentiment analysis in multilingual Web texts." In: *Information retrieval* 5, pp. 526–558.

Borge-Holthoefer, Javier, Alejandro Rivero, Iñigo García, Elisa Cauhé, Alfredo Ferrer, Darío Ferrer, David Francos, David Iniguez, María Pilar Pérez, Gonzalo Ruiz, et al. (2011). "Structural and dynam-

ical patterns on online social networks: the Spanish May 15th movement as a case study." In: *PloS one* 6.8, e23883.

Borondo, J, A J Morales, J C Losada, and R M Benito (2012). "Characterizing and modeling an electoral campaign in the context of Twitter: 2011 Spanish presidential election as a case study." In: *AIP Chaos* 22.2, p. 23138.

Bosco, C., M. Lai, V. Patti, F. M. Rangel Pardo, and P Rosso (2016). "Tweeting in the Debate about Catalan Elections." In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016). Emotion and Sentiment Analysis Workshop.* Pp. 67–70.

Bradley, Margaret M and Peter J Lang (1999). *Affective norms for English words (ANEW): Instruction manual and affective ratings*. Tech. rep. Technical report C-1, the center for research in psychophysiology, University of Florida.

Brants, Thorsten (2000). "TnT: a statistical part-of-speech tagger." In: *Proceedings of the sixth conference on Applied natural language processing*. Association for Computational Linguistics, pp. 224–231.

Brill, E (1992). "A simple rule-based part of speech tagger." In: *Proceedings of the workshop on Speech and Natural Language*. HLT'91. Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 112–116. ISBN: 1-55860-272-0. DOI: 10.3115/1075527.1075553.

Broersma, Marcel and Todd Graham (2012). "Social media as beat: Tweets as a news source during the 2010 British and Dutch elections." In: *Journalism Practice* 6.3, pp. 403–419.

Brooke, J, M Tofiloski, and M Taboada (2009). "Cross-Linguistic Sentiment Analysis: From English to Spanish." In: *Proceedings of the International Conference RANLP-2009*. Borovets, Bulgaria: ACL, pp. 50–54.

Brown, P and S Levinson (1987). *Politeness, Some universals in language use*. Cambridge, Cambridge University Press.

Buchholz, S and E Marsi (2006). "CoNLL-X shared task on multilingual dependency parsing." In: *Proceedings of the Tenth Conference on Computational Natural Language Learning*. Association for Computational Linguistics, pp. 149–164.

Buckley, Chris and Ellen M Voorhees (2000). "Evaluating evaluation measure stability." In: *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, pp. 33–40.

CIS (2014). *Barómetro de Octubre 2014*. Technical Report 3041. Centro de Investigaciones Sociológicas.

Caldarelli, Guido, Alessandro Chessa, Fabio Pammolli, Gabriele Pompa, Michelangelo Puliga, Massimo Riccaboni, and Gianni Riotta (2014). "A multi-level geographical study of Italian political elections from Twitter data." In: *PloS one* 9.5, e95809.

Cambria, Erik, Daniel Olsher, and Rajagopal Dheeraj (2014). "SenticNet 3: a common and common-sense knowledge base for cognition-driven sentiment analysis." In: *Twenty-eighth AAAI conference on artificial intelligence*, pp. 1515–1521.

Cambria, Erik, Amir Hussain, Catherine Havasi, and Chris Eckl (2009). "AffectiveSpace: blending common sense and affective knowledge to perform emotive reasoning." In: *WOMSA at CAEPIA, Seville*, pp. 32–41.

Campos, Héctor (1993). *De la oración simple a la oración compuesta: curso superior de gramática española*. Georgetown University Press.

Carbonell, Jaime Guillermo (1979). *Subjective Understanding: Computer Models of Belief Systems.* Tech. rep. DTIC Document.

Carvalho, P., L. Sarmento, J. Teixeira, and M. J. Silva (2011). "Liars and saviors in a sentiment annotated corpus of comments to political debates." In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*. Association for Computational Linguistics, pp. 564–568.

Cavnar, William B and John M Trenkle (1994). "N-Gram-Based Text Categorization." In: *Proceedings of Third Annual Symposium on Document Analysis and Information Retrieval*. Las Vegas, Nevada, USA: UNLV Publications/Reprographics, pp. 161–175.

Ceron, Andrea, Luigi Curini, and Stefano M Iacus (2015a). "Using Sentiment Analysis to Monitor Electoral Campaigns: Method Matters—Evidence From the United States and Italy." In: *Social Science Computer Review* 33.1, pp. 3–20.

— (2015b). "Using social media to forecast electoral results. A meta-analysis." In: *Italian Journal of Applied Statistics*.

Chang, C. and C. Lin (2011). "LIBSVM: A library for support vector machines." In: *ACM Transactions on Intelligent Systems Technology* 2.3, 27:1–27:27. ISSN: 2157-6904. DOI: 10.1145/1961189.1961199.

Chen, Boxing and Xiaodan Zhu (2014). "Bilingual Sentiment Consistency for Statistical Machine Translation." In: *The 52nd Annual Meeting of the Association for Computational Linguistics. Proceedings of the Conference. Volume 1: Long Papers. ACL 2014*. ACL. Baltimore, pp. 607–615.

Chen, Danqi and Christopher D Manning (2014). "A Fast and Accurate Dependency Parser using Neural Networks." In: *EMNLP*, pp. 740–750.

Chen, W., Y. Wu, and H. Isahara (2008). "Learning Reliable Information for Dependency Parsing Adaptation." In: *Proceedings of the 22nd International Conference on Computational Linguistics - Volume 1*. COLING '08. Manchester, United Kingdom: Association for Computational Linguistics, pp. 113–120. ISBN: 978-1-905593-44-6.

Chen, Yanqing and Steven Skiena (2014). "Building Sentiment Lexicons for All Major Languages." In: *The 52nd Annual Meeting of the*

Association for Computational Linguistics. Proceedings of the Conference. Volume 2: Short Papers. ACL 2014. ACL. Baltimore, pp. 383–389.

Cheng, Alex and Oles Zhulyn (2012). "A System for Multilingual Sentiment Learning On Large Data Sets." In: *COLING*, pp. 577–592.

Chowdhury, Md. Faisal Mahbub, Marco Guerini, Sara Tonelli, and Alberto Lavelli (2013). "FBK: Sentiment Analysis in Twitter with Tweetsted." In: *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Seventh International Workshop on Semantic Evaluation (SemEval 2013)*. ACL. Atlanta, Georgia, pp. 466–470.

Collins, Michael (1997). "Three generative, lexicalised models for statistical parsing." In: *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, pp. 16–23.

Congosto, María Luz, Montse Fernández, and Esteban Moro (2011). "Twitter y política: Información, opinión y¿Predicción?" In:

Conover, Michael D, Bruno Gonçalves, Alessandro Flammini, and Filippo Menczer (2012). "Partisan asymmetries in online political activity." In: *EPJ Data Science* 1.1, Article: 6.

Conover, Michael, Jacob Ratkiewicz, Matthew R Francisco, Bruno Gonçalves, Filippo Menczer, and Alessandro Flammini (2011). "Political Polarization on Twitter." In: *ICWSM* 133, pp. 89–96.

Contractor, Danish and Tanveer A Faruquie (2013). "Understanding Election Candidate Approval Ratings Using Social Media Data." In: *WWW 2013 Companion*. Rio de Janeiro: ACM Press, pp. 189–190.

Covington, Michael A (2001). "A fundamental algorithm for dependency parsing." In: *Proceedings of the 39th annual ACM southeast conference*, pp. 95–102.

Criado, J Ignacio, Guadalupe Martínez-Fuentes, and Aitor Silván (2013). "Twitter en campaña: las elecciones municipales españolas de 2011." In: *Revista de investigaciones Políticas y Sociológicas* 12.1, pp. 93–113.

Cruz Mata, F. L. (2011). "Extracción de opiniones sobre características: Un enfoque práctico adaptado al dominio." PhD thesis. Universidad de Sevilla.

Cruz, F. L, J. A Troyano, B. Pontes, and F. J. Ortega (2014a). "ML-SentiCon: Un lexicón multilingüe de polaridades semánticas a nivel de lemas." In: *Procesamiento del Lenguaje Natural* 53, pp. 113–120.

Cruz, Fermín L, José A Troyano, Beatriz Pontes, and F Javier Ortega (2014b). "Building layered, multilingual sentiment lexicons at synset and lemma levels." In: *Expert Systems with Applications* 41.13, pp. 5984–5994.

Cruz, Noa P, Maite Taboada, and Ruslan Mitkov (2015). "A machine-learning approach to negation and speculation detection for sentiment analysis." In: *Journal of the Association for Information Science and Technology*.

Cui, Anqi, Min Zhang, Yiqun Liu, and Shaoping Ma (2011). "Emotion Tokens: Bridging the Gap Among Multilingual Twitter Sentiment Analysis." In: *Information Retrieval Technology. 7th Asia Information Retrieval Societies Conference, AIRS 2011, Dubai, United Arab Emirates, December 18-20, 2011. Proceedings*. Ed. by Mohamed Vall Mohamed Salem, Khaled Shaalan, Farhad Oroumchian, Azadeh Shakery, and Halim Khelalfa. Vol. 7097. Lecture Notes in Computer Science. Berlin and Heidelberg: Springer, pp. 238–249.

Davies, Alex and Zoubin Ghahramani (2011). "Language-independent Bayesian sentiment mining of Twitter." In: *The 5th SNA-KDD Workshop'11 (SNA-KDD'11)*. ACM. San Diego, CA.

De Marneffe, Marie-Catherine and Christopher D Manning (2008). "The Stanford typed dependencies representation." In: *Coling 2008: Proceedings of the workshop on Cross-Framework and Cross-Domain Parser Evaluation*. Association for Computational Linguistics, pp. 1–8.

Deltell, Luis (2012). "Estrategias de comunicación política en las redes sociales durante la campaña electoral del 2011 en España: el caso de eQuo." In: *II Jornadas de la Asociación Madrileña de Sociología*. Madrid.

Deltell, Luis, Florencia Claes, and José Miguel Osteso (2013). "Predicción de tendencia política por Twitter: Elecciones Andaluzas 2012." In: *Ambitos: Revista internacional de comunicación* 22, pp. 91–100.

Demirtas, Erkin and Mykola Pechenizkiy (2013). "Cross-lingual polarity detection with machine translation." In: *Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining*. ACM, p. 9.

Derks, Daantje, Agneta H Fischer, and Arjan ER Bos (2008). "The role of emotion in computer-mediated communication: A review." In: *Computers in Human Behavior* 24.3, pp. 766–785.

DiGrazia, Joseph, Karissa McKelvey, Johan Bollen, and Fabio Rojas (2013). "More Tweets, More Votes: Social Media as a Quantitative Indicator of Political behavior." In: *PLOS ONE* 8.11, e79449.

Dyer, Chris, Miguel Ballesteros, Wang Ling, Austin Matthews, and Noah A. Smith (2015). "Transition-Based Dependency Parsing with Stack Long Short-Term Memory." In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Beijing, China: Association for Computational Linguistics, pp. 334–343.

Džeroski, Sašo, Tomaž Erjavec, Nina Ledinek, Petr Pajas, Zdenek Žabokrtsky, and Andreja Žele (2006). "Towards a Slovene dependency

treebank." In: *Proc. of the Fifth Intern. Conf. on Language Resources and Evaluation (LREC)*.

Effing, Robin, Jos van Hillegersberg, and Theo Huibers (2011). "Social Media and Political Participation: Are Facebook, Twitter and YouTube Democratizing Our Political Systems?" In: *Electronic Participation*. Ed. by Efthimios Tambouris, Ann Macintosh, and Hans de Bruijn. Vol. 6847. Lecture Notes in Computer Science. Berlin and Heidelberg: Springer, pp. 25–35.

Esuli, Andrea and Fabrizio Sebastiani (2006). "Sentiwordnet: A publicly available lexical resource for opinion mining." In: *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*. Vol. 6. Genoa, Italy, pp. 417–422.

— (2015). "Optimizing Text Quantifiers for Multivariate Loss Functions." In: *ACM Trans. Knowl. Discov. Data* 9.4, 27:1–27:27. ISSN: 1556-4681. DOI: 10.1145/2700406.

Fan, Rong-En, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin (2008). "LIBLINEAR: A library for large linear classification." In: *The Journal of Machine Learning Research* 9, pp. 1871–1874.

Fink, Clay, Nathan Bos, Alexander Perrone, Edwina Liu, and Jonathon Kopecky (2013). "Twitter, Public Opinion, and the 2011 Nigerian Presidential Election." In: *Proceedings of SocialCom/PASSAT/BigData/EconCom/BioMedCom 2013 (SocialCom 2013)*. IEEE Computer Society. Washington, D.C., USA, pp. 311–320.

Forcada, M. L, M. Ginestí-Rosell, J. Nordfalk, J. O'Regan, S. Ortiz-Rojas, J. A. Pérez-Ortiz, F. Sánchez-Martínez, G. Ramírez-Sánchez, and F. M. Tyers (2011). "Apertium: a free/open-source platform for rule-based machine translation." In: *Machine translation* 25.2, pp. 127–144.

Foster, Jennifer, Özlem Çetino\uglu, Joachim Wagner, Joseph Le Roux, Stephen Hogan, Joakim Nivre, Deirdre Hogan, and Josef van Genabith (2011). "#hardtoparse: POS Tagging and Parsing the Twitterverse." In: *The AAAI-11 Workshop on Analyzing Microtext*. AAAI. San Francisco, CA.

Gamallo, P., M. García, and S. Fernández Lanza (2013). "TASS: A Naive-Bayes strategy for sentiment analysis on Spanish tweets." In: *XXIX Congreso de la Sociedad Española de Procesamiento de Lenguaje Natural (SEPLN 2013). TASS 2013 - Workshop on Sentiment Analysis at SEPLN 2013*. Ed. by Alberto Díaz Esteban, Iñaki Alegría Loinaz, and Julio Villena Román. Madrid, Spain, pp. 126–132.

Gamon, Michael (2004). "Sentiment classification on customer feedback data: noisy data, large feature vectors, and the role of linguistic analysis." In: *Proceedings of the 20th international conference on Computational Linguistics*. Association for Computational Linguistics, p. 841.

Gao, Dehong, Furu Wei, Wenjie Li, Xiaohua Liu, and Ming Zhou (2013). "Cotraining Based Bilingual Sentiment Lexicon Learning." In: *Workshops at the Twenty-Seventh AAAI Conference on Artificial Intelligence. AAAI Conference Late-Breaking Papers*. AAAI. Bellevue, Washington, USA.

García, David and Mike Thelwall (2013). "Political alignment and emotional expression in Spanish Tweets." In: *Proceedings of the TASS workshop at SEPLN*, pp. 151–159.

Garcia, M. and P. Gamallo (2015). "Yet Another Suite of Multilingual NLP Tools." In: *Languages, Applications and Technologies. Communications in Computer and Information Science*. Vol. 563. Springer, pp. 65–75.

Gaurav, Manish, Amit Srivastava, Anoop Kumar, and Scott Miller (2013). "Leveraging Candidate Popularity on Twitter to Predict Election Outcome." In: *Proceedings of the 7th Workshop on Social Network Mining and Analysis (SNA-KDD 2013)*. ACM. Chicago, IL, Article No. 7.

Gayo-Avello, Daniel (2011). "Don't Turn Social Median Into Another 'Literay Digest' Poll." In: *Communications of the ACM* 54.10, pp. 121–128.

— (2012). "No, you cannot predict elections with Twitter." In: *IEEE Internet Computing* 16.6, pp. 91–94.

Gayo-Avello, Daniel, Panagiotis T Metaxas, and Eni Mustafaraj (2011). "Limits of Electoral Predictions Using Twitter." In: *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media (ICWSM 2011)*. AAAI. Barcelona, Spain, pp. 490–493.

Ghorbel, Hatem and David Jacot (2011). "Sentiment analysis of French movie reviews." In: *Advances in Distributed Agent-Based Retrieval Tools*. Springer, pp. 97–108.

Gil, Iván (2014). *Auge y caída de Tania Sánchez: de ganar las primarias de IU a peligrar su candidatura*. Ed. by www.elconfidencial.com.

Giménez, Jesús and Lluis Marquez (2004). "Fast and accurate part-of-speech tagging: The SVM approach revisited." In: *Recent Advances in Natural Language Processing III*, pp. 153–162.

Gimpel, K, N Schneider, B O'connor, D Das, D Mills, J Eisenstein, M Heilman, D Yogatama, J Flanigan, and N A Smith (2011). "Part-of-speech tagging for Twitter: annotation, features, and experiments." In: *HLT '11 Proc. of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers* 2, pp. 42–47.

Go, Alec, Richa Bhayani, and Lei Huang (2009). "Twitter sentiment classification using distant supervision." In: *CS224N Project Report, Stanford* 1.12.

Golbeck, Jennifer and Derek L Hansen (2011). "Computing Political Preference among Twitter Followers." In: *Proceedings of the*

SIGCHI Conference on Human Factors in Computing Systems (CHI 2011). ACM. Vancouver, BC, Canada, pp. 1105–1108.

Gómez-Rodríguez, Carlos and Joakim Nivre (2010). "A transition-based parser for 2-planar dependency structures." In: *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, pp. 1492–1501.

Gómez-Rodríguez, Carlos, Francesco Sartorio, and Giorgio Satta (2014). "A polynomial-time dynamic oracle for non-projective dependency parsing." In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 917–927.

Goutam, R. and B. R. Ambati (2011). "Exploring self training for Hindi dependency parsing." In: *Proceedings of 5th International Joint Conference on Natural Language Processing*. Chiang Mai, Thailand: Asian Federation of Natural Language Processing, pp. 1452–1456.

Granado, Olga (2014). *Íñigo Errejón, suspendido de empleo y sueldo por la Universidad de Málaga*. Ed. by www.eldiario.es.

Grant, Will J, Brenda Moon, and Janie Busby Grant (2010). "Digital Dialogue? Australian Politicians' use of the Social Network Tool Twitter." In: *Australian Journal of Political Science* 45.4, pp. 579–604.

Greene, Stephen and Philip Resnik (2009). "More than Words: Syntactic Packaging and Implicit Sentiment." In: *NAACL'09 Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. ACL. Boulder, Colorado, pp. 503–511.

Gui, Lin, Ruifeng Xu, Jun Xu, Li Yuan, Yuanlin Yao, Jiyun Zhou, Qiaoyun Qiu, Shuwei Wang, Kam-Fai Wong, and Ricky Cheung (2013). "A Mixed Model for Cross Lingual Opinion Analysis." In: *Natural Language Processing and Chinese Computing*. Ed. by Guodong Zhou, Juanzi Li, Dongyan Zhao, and Yansong Feng. Vol. 400. Communications in Computer and Information Science. Heidelberg, NewYork, Dordrecht and London: Springer, pp. 93–104.

Gui, Lin, Ruifeng Xu, Qin Lu, Jun Xu, Jiang Xu, Bin Liu, and Xiaolong Wang (2014). "Cross-lingual Opinion Analysis via Negative Transfer Detection." In: *The 52nd Annual Meeting of the Association for Computational Linguistics. Proceedings of the Conference. Volume 2: Short Papers. ACL 2014*. ACL. Baltimore, pp. 860–865.

Guo, Jiang, Wanxiang Che, Haifeng Wang, and Ting Liu (2016). "Exploiting Multi-typed Treebanks for Parsing with Deep Multi-task Learning." In: *arXiv preprint arXiv:1606.01161*.

Habernal, Ivan, Tomáš Ptáček, and Josef Steinberger (2014). "Supervised sentiment analysis in Czech social media." In: *Information Processing & Management* 50.5, pp. 693–707.

Hajmohammadi, Mohammad Sadegh, Roliana Ibrahim, and Ali Selamat (2014). "Bi-view semi-supervised active learning for cross-

lingual sentiment classification." In: *Information Processing and Management* 50.5, pp. 718–732.

Hall, Mark, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten (2009). "The WEKA data mining software: an update." In: *ACM SIGKDD explorations newsletter* 11.1, pp. 10–18.

Harfoush, Rahaf (2009). *Yes We Did! An inside look at how social media built the Obama brand*. New Riders.

Hayes, A.F. and K. Krippendorff (2007). "Answering the call for a standard reliability measure for coding data." In: *Communication Methods and Measures* 1.1, pp. 77–89.

Hiroshi, Kanayama, Nasukawa Tetsuya, and Watanabe Hideo (2004). "Deeper sentiment analysis using machine translation technology." In: *Proceedings of the 20th international conference on Computational Linguistics (COLING 2004)*. Geneva, Switzerland.

Hogenboom, Alexander, Bas Heerschop, Flavius Frasincar, Uzay Kaymak, and Franciska de Jong (2014). "Multi-lingual support for lexicon-based sentiment analysis guided by semantics." In: *Decision support systems* 62, pp. 43–53.

Howard, Philip N, Aiden Duffy, Deen Freelon, Muzammil M Hussain, Will Mari, and Marwa Maziad (2011). *Opening closed regimes: what was the role of social media during the Arab Spring?* Tech. rep.

Hu, Minqing and Bing Liu (2004). "Mining and summarizing customer reviews." In: *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, pp. 168–177.

Huberty, Mark (2013). "Multi-Cycle Forecasting of Congressional Elections with Social Media." In: *Proceedings of the 2nd workshop on Politics, elections and data (PLEAD 2013)*. ACM. San Francisco, CA, pp. 23–29.

Hurtado, L. F., F. Pla, and D. Buscaldi (2015). "ELiRF-UPV en TASS 2015: Análisis de Sentimientos en Twitter." In: *Proceedings of TASS 2015: Workshop on Sentiment Analysis at SEPLN*, pp. 35–40.

Inrak, Piyatida and Sukree Sinthupinyo (2010). "Applying latent semantic analysis to classify emotions in Thai text." In: *Computer Engineering and Technology (ICCET), 2010 2nd International Conference on*. Vol. 6. IEEE, pp. V6–450.

Jensen, Michael J. and Nick Anstead (2013). "Psephological investigations: Tweets, votes, and unknown unknowns in the republican nomination process." In: *Policy and Internet* 5.2, pp. 161–182. ISSN: 19442866. DOI: 10.1002/1944-2866.POI329.

Jia, Lifeng, Clement Yu, and Weiyi Meng (2009). "The effect of negation on Sentiment Analysis and Retrieval Effectiveness." In: *CIKM'09 Proceeding of the 18th ACM conference on Information and knowledge management*. ACM. Hong Kong: ACM Press, pp. 1827–1830.

Joshi, M and C Penstein-Rosé (2009). "Generalizing dependency features for opinion mining." In: *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*. ACLShort '09. Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 313–316.

Jungherr, A, P Jürgens, and H Schoen (2012). "Why the Pirate Party Won the German Election of 2009 or The Trouble With Predictions: A Response to Tumasjan, A., Sprenger, T. O., Sander, P. G., & Welpe, I. M. Predicting Elections With Twitter: What 140 Characters Reveal About Political Sentime." In: *Social Science Computer Review,* 30.2, pp. 229–234.

Jurafsky, D. and Martin J.H (2016). *Classification: Naive Bayes, Logistic Regression, Sentiment. Chapter 7 of Speech and Language Processing (3rd ed. draft)*.

Kalchbrenner, Nal, Edward Grefenstette, and Phil Blunsom (2014). "A Convolutional Neural Network for Modelling Sentences." In: *The 52nd Annual Meeting of the Association for Computational Linguistics. Proceedings of the Conference. Volume 1: Long Papers*. ACL. Baltimore, Maryland, USA, pp. 655–665.

Kanayama, Hiroshi and Tetsuya Nasukawa (2006). "Fully automatic lexicon expansion for domain-oriented sentiment analysis." In: *Proceedings of the 2006 conference on empirical methods in natural language processing*. Association for Computational Linguistics, pp. 355–363.

Kennedy, Alistair and Diana Inkpen (2006). "Sentiment classification of movie reviews using contextual valence shifters." In: *Computational intelligence* 22.2, pp. 110–125.

Kim, Jungi, Hun-Young Jung, Sang-Hyob Nam, Yeha Lee, and Jong-Hyeok Lee (2009). "Found in translation: conveying subjectivity of a lexicon of one language into another using a bilingual dictionary and a link analysis algorithm." In: *International Conference on Computer Processing of Oriental Languages*. Springer, pp. 112–121.

Kiritchenko, Svetlana, Xiaodan Zhu, and Saif M Mohammad (2014). "Sentiment Analysis of Short Informal Texts." In: *Journal of Artificial Intelligence Research* 50.1, pp. 723–762.

Klinger, Roman and Philipp Cimiano (2014). "The USAGE review corpus for fine-grained, multi-lingual opinion analysis." In: *Proceedings of the Language Resources and Evaluation Conference*.

LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton (2015). "Deep learning." In: *Nature* 521.7553, pp. 436–444.

Li, Yung-Ming and Tsung-Ying Li (2013). "Deriving market intelligence from microblogs." In: *Decision Support Systems* 55.1, pp. 206–217.

Liu, Hugo and Push Singh (2004). "ConceptNet—a practical commonsense reasoning tool-kit." In: *BT technology journal* 22.4, pp. 211–226.

Liu, Qian, Zhiqiang Gao, Bing Liu, and Yuanlin Zhang (2016). "Automated rule selection for opinion target extraction." In: *Konwledge-Based Systems* 104, pp. 74–88.

Livne, Avishay, Matthew P Simmons, Eytan Adar, and Lada A Adamic (2011). "The Party is Over Here: Structure and Content in the 2010 Election." In: *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media (ICWSM 2011)*. AAAI. Barcelona, Spain, pp. 201–208.

Lui, Marco and Timothy Baldwin (2012). "langid. py: An off-the-shelf language identification tool." In: *Proceedings of the ACL 2012 system demonstrations*. Association for Computational Linguistics, pp. 25–30.

Lynn, Teresa, Jennifer Foster, Mark Dras, and Lamia Tounsi (2014). "Cross-lingual Transfer Parsing for Low-Resourced Languages: An Irish Case Study." In: *CLTW 2014. The First Celtic Language Technology Workshop. Proceedings of the Workshop*. Dublin, Ireland, pp. 41–49.

Makazhanov, Aibek and Davood Rafiei (2013). "Predicting Political Preference of Twitter Users." In: *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2013)*. ACM. Niagara, ON, Canada, pp. 298–305.

Marchetti-Bowick, Micol and Nathanel Chambers (2012). "Learning for Microblogs with Distant Supervision: Political Forecasting with Twitter." In: *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*. ACL. Avignon, France, pp. 603–612.

Martínez Cámara, Eugenio, Martín Valdivia, María Teresa, José Manuel Perea Ortega, and Luis Alfonso Ureña López (2011). "Técnicas de clasificación de opiniones aplicadas a un corpus en español." In: *Procesamiento del Lenguaje Natural*.

Martínez Cámara, Eugenio, M Teresa Martín Valdivia, M Dolores Molina-González, and José M Perea-Ortega (2014). "Integrating Spanish lexical resources by meta-classifiers for polarity classification." In: *Journal of Information Science* 40.4, pp. 538–554.

Martins, Andre, Miguel Almeida, and Noah A. Smith (2013). "Turning on the Turbo: Fast Third-Order Non-Projective Turbo Parsers." In: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Sofia, Bulgaria, pp. 617–622.

McDonald, R et al. (2013). "Universal Dependency Annotation for Multilingual Parsing." In: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, pp. 92–97. ISBN: 9781937284510.

McDonald, Ryan, Fernando Pereira, Kiril Ribarov, and Jan Hajič (2005). "Non-projective dependency parsing using spanning tree algo-

rithms." In: *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pp. 523–530.

McKnight, Patrick E and Julius Najab (2010). "Mann-Whitney U Test." In: *Corsini Encyclopedia of Psychology*.

Medagoda, Nishantha, Subana Shanmuganathan, and Jacqueline Whalley (2013). "A comparative analysis of opinion mining and sentiment classification in non-English languages." In: *Advances in ICT for Emerging Regions (ICTer), 2013 International Conference on*. IEEE, pp. 144–148.

Metaxas, Panagiotis T, Eni Mustafaraj, and Dani Gayo-Avello (2011). "How (not) to predict elections." In: *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on*. IEEE, pp. 165–171.

Miller, George A (1995). "WordNet: a lexical database for English." In: *Communications of the ACM* 38.11, pp. 39–41.

Mitchell, Tom M (1997). "Machine learning. 1997." In: *Burr Ridge, IL: McGraw Hill* 45, p. 37.

Mohammad, Saif M, Svetlana Kiritchenko, and Xiaodan Zhu (2013). "NRC-Canada: Building the State-of-the-Art in Sentiment Analysis of Tweets." In: *Proceedings of the seventh international workshop on Semantic Evaluation Exercises (SemEval-2013)*. Atlanta, Georgia, USA.

Momtazi, Saeedeh (2012). "Fine-grained German Sentiment Analysis on Social Media." In: *8th International Conference on Language Resources and Evaluation (LREC'12)*, pp. 1215–1220.

Montejo-Ráez, A, E Martínez-Cámara, M T Martín-Valdivia, and L A Ureña-López (2012). "Random Walk Weighting over SentiWordNet for Sentiment Polarity Detection on Twitter." In: *WASSA 2012, 3rdWorkshop on Computational Approaches to Subjectivity and Sentiment Analysis, Proceedings of the Workshop*. Jeju, Republic of Korea, pp. 3–10.

Monti, Corrado, Alessandro Rozza, Giovanni Zappella, Matteo Zignani, Adam Arvidsson, and Elanor Colleoni (2013). "Modelling political disaffection from Twitter data." In: *Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining*. ACM, p. 3.

Montoyo, Andrés, Patricio Martínez-Barco, and Alexandra Balahur (2012). "Subjectivity and sentiment analysis: An overview of the current state of the art and envisaged developments." In: *Decision Support Systems* 53.4, pp. 675–679.

Nakagawa, Tetsuji, Kentaro Inui, and Sadao Kurohashi (2010). "Dependency Tree-Based Sentiment Classification using CRFs with Hidden Variables." In: *NAACL HLT'10 Human Language Technologies: The 2010 Annual Conference of the North American Chapter of*

the Association for Computational Linguistics. Proceedings of the Main Conference. ACL. Los Angeles, CA, pp. 786–794.

Nakov, P., S. Rosenthal, Z. Kozareva, V. Stoyanov, A. Ritter, and T. Wilson (2013). "SemEval-2013 Task 2: Sentiment Analysis in Twitter." In: *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Seventh International Workshop on Semantic Evaluation (SemEval 2013)*. ACL. Atlanta, Georgia, pp. 312–320.

Nakov, Preslav, Alan Ritter, Sara Rosenthal, Fabrizio Sebastiani, and Veselin Stoyanov (2016a). "SemEval-2016 task 4: Sentiment analysis in Twitter." In: *Proceedings of the 10th international workshop on semantic evaluation (SemEval 2016), San Diego, US*, pp. 1–18.

— (2016b). "SemEval-2016 task 4: Sentiment analysis in Twitter." In: *Proceedings of the 10th international workshop on semantic evaluation (SemEval 2016), San Diego, US*.

Narr, Sascha, Michael Hülfenhaus, and Sahin Alnayrak (2012). "Language-Independent Twitter Sentiment Analysis." In: *Workshop on Knowledge Discovery, Data Mining and Machine Learning (KDML-2012)*. Dortmund, Germany.

Neri, Federico, Carlo Aliprandi, Federico Capeci, Montserrat Cuadros, and Tomas By (2012). "Sentiment analysis on social media." In: *Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012)*. IEEE Computer Society, pp. 919–926.

Nivre, J. (2008a). "Algorithms for Deterministic Incremental Dependency Parsing." In: *Computational Linguistics* 34.4, pp. 513–553. ISSN: 0891-2017. DOI: 10.1162/coli.07-056-R1-07-027.

Nivre, Joakim (2003). "An efficient algorithm for projective dependency parsing." In: *Proceedings of the 8th International Workshop on Parsing Technologies (IWPT*.

— (2008b). "Algorithms for deterministic incremental dependency parsing." In: *Computational Linguistics* 34.4, pp. 513–553.

Nivre, Joakim, Johan Hall, Jens Nilsson, Atanas Chanev, Gülsen Eryigit, Sandra Kübler, Svetoslav Marinov, and Erwin Marsi (2007). "MaltParser: A language-independent system for data-driven dependency parsing." In: *Natural Language Engineering* 13.2, pp. 95–135.

Nivre, Joakim, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, et al. (2016). "Universal dependencies v1: A multilingual treebank collection." In: *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, pp. 1659–1666.

O'Connor, Brendan, Ramnath Balasubramanyan, Bryan Routledge, and Noah A Smith (2010). "From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series." In: *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*

(ICWSM 2010). Ed. by William W Cohen and Samuel Gosling. AAAI. Washington, DC, pp. 122–129.

Oswald, Hans and Christine Schmid (1998). "Political participation of young people in East Germany." In: *German Politics* 7.3, pp. 147–164.

Padró, L. and E. Stanilovsky (2012). "Freeling 3.0: Towards wider multilinguality." In: *Proceedings of the 8th edition of the Language Resources and Evaluation Conference (LREC 2012)*. Istambul.

Pak, A and P Paroubek (2010). "Twitter as a Corpus for Sentiment Analysis and Opinion Mining." In: *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*. Valletta, Malta: European Language Resources Association (ELRA). ISBN: 2-9517408-6-7.

Paltoglou, Georgios and Mike Thelwall (2010). "A study of Information Retrieval weighting schemes for sentiment analysis." In: *ACL2010. The 48th Annual Meeting of the Association for Computational Linguistics. Conference Proceedings*. ACL. Uppsala, Sweden, pp. 1386–1395.

Pang, B. and L. Lee (2005). "Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales." In: *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, pp. 115–124.

Pang, B and L Lee (2008). *Opinion Mining and Sentiment Analysis*. Hanover, MA, USA: now Publishers Inc.

Pang, B, L Lee, and S Vaithyanathan (2002). "Thumbs up? Sentiment classification using machine learning techniques." In: *Proceedings o f EMNLP*, pp. 79–86.

Pang, Bo and Lillian Lee (2004). "A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts." In: *Proceedings of the 42nd annual meeting on Association for Computational Linguistics*. Association for Computational Linguistics, pp. 271–278.

Pennebaker, J W, M E Francis, and R J Booth (2001). "Linguistic inquiry and word count: LIWC 2001." In: *Mahway: Lawrence Erlbaum Associates*, p. 71.

Pennington, Jeffrey, Richard Socher, and Christopher D Manning (2014). "Glove: Global Vectors for Word Representation." In: *EMNLP*. Vol. 14, pp. 1532–1543.

Perea-Ortega, José M, M Teresa Martín-Valdivia, L Alfonso Ureña-López, and Eugenio Martínez-Cámara (2013). "Improving Polarity Classification of Bilingual Parallel Corpora Combining Machine Learning and Semantic Orientation Approaches." In: *Journal of the American Society for Information Science and Technology* 64.9, pp. 1864–1877.

Petrov, S, D Das, and R McDonald (2011). "A universal part-of-speech tagset." In: *arXiv preprint arXiv:1104.2086*.

Pew-Research-Center (2011). *Litle Change in Public's Response to 'Capitalism', 'Socialism'*. Pew Research Center.

Pla, Ferran and Lluís-F. Hurtado (2013). "ELiRF-UPV en TASS-2013: Análisis de Sentimientos en Twitter." In: *XXIX Congreso de la Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN 2013). TASS 2013 - Workshop on Sentiment Analysis at SEPLN 2013*. Ed. by Alberto Díaz Esteban, Iñaki Alegría Loinaz, and Julio Villena Román. SEPLN. Madrid, Spain, pp. 220–227.

Plank, Barbara, Anders Søgaard, and Yoav Goldberg (2016). "Multilingual Part-of-Speech Tagging with Bidirectional Long Short-Term Memory Models and Auxiliary Loss." In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Berlin, Germany: Association for Computational Linguistics, pp. 412–418.

Platt, J C (1999). "Advances in kernel methods." In: ed. by Bernhard Schölkopf, Christopher J C Burges, and Alexander J Smola. Cambridge, MA, USA: MIT Press. Chap. Fast train, pp. 185–208. ISBN: 0-262-19416-3.

Pontiki, Maria, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar (2014). "Semeval-2014 task 4: Aspect based sentiment analysis." In: *Proceedings of SemEval*, pp. 27–35.

Poria, Soujanya, Erik Cambria, and Alexander Gelbukh (2015). "Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis." In: *Proceedings of EMNLP*, pp. 2539–2544.

— (2016). "Aspect extraction for opinion mining with a deep convolutional neural network." In: *Knowledge-Based Systems* 108, pp. 42–49.

Poria, Soujanya, Erik Cambria, Grégoire Winterstein, and Guang-Bin Huang (2014). "Sentic patterns: Dependency-based rules for concept-level sentiment analysis." In: *Knowledge-Based Systems* 69, pp. 45–63.

RTVE.es/EFE (2014). *La Universidad de Málaga suspende cautelarmente de empleo y sueldo a Íñigo Errejón*. Ed. by www.rtve.es.

Ramírez-Esparza, N, J W Pennebaker, F A García, and R Suria (2007). "La psicología del uso de las palabras: Un programa de computadora que analiza textos en Español (The psychology of word use: A computer program that analyzes texts in Spanish)." In: *Revista Mexicana de Psicología* 24, pp. 85–99.

Rasooli, Mohammad Sadegh and Joel R. Tetreault (2015). "Yara Parser: A Fast and Accurate Dependency Parser." In: *CoRR* abs/1503.06733.

Ratnaparkhi, Adwait et al. (1996). "A maximum entropy model for part-of-speech tagging." In: *Proceedings of the conference on em-*

pirical methods in natural language processing. Vol. 1. Philadelphia, USA, pp. 133–142.

Razzaq, Muhammad Asif, Ali Mustafa Qamar, and Hafiz Syed Muhammad Bilal (2014). "Prediction and analysis of Pakistan election 2013 based on sentiment analysis." In: *Advances in Social Networks Analysis and Mining (ASONAM), 2014 IEEE/ACM International Conference on*. IEEE, pp. 700–703.

Redondo, Jaime, Isabel Fraga, Isabel Padrón, and Montserrat Comesaña (2007). "The Spanish adaptation of ANEW (affective norms for English words)." In: *Behavior research methods* 39.3, pp. 600–605.

Riveiro, Aitor (2014). *Tania Sánchez reclama una nueva dirección de IU de Madrid que refleje su triunfo en las primarias*. Ed. by www.eldiario.es.

Román, J., E. Martínez-Cámara, J. García-Morera, and Salud M. Jiménez-Zafra (2015). "TASS 2014-The Challenge of Aspect-based Sentiment Analysis." In: *Procesamiento del Lenguaje Natural* 54, pp. 61–68.

Romero-Frías, Esteban and Liwen Vaughan (2012). "Exploring the Relationships Between Media and Political Parties Through web Hyperlink Analysis: The Case of Spain." In: *Journal of the American Society for Information Science and Technology* 63.5, pp. 967–976.

Rosenthal, S., P. Nakov, A. Ritter, and V. Stoyanov (2014). "Semeval-2014 task 9: Sentiment analysis in Twitter." In: *Proceedings of The 8th InternationalWorkshop on Semantic Evaluation (SemEval 2014)*, pp. 411–415.

Rui, Huaxia, Yizao Liu, and Andrew Whinston (2013). "Whose and what chatter matters? The effect of tweets on movie sales." In: *Decision Support Systems* 55.4, pp. 863–870.

Salton, Gerard and Michael J McGill (1986). *Introduction to modern information retrieval*. McGraw-Hill, Inc.

San Vicente, I. and X. Saralegi (2016). "Polarity Lexicon Building: to what Extent Is the Manual Effort Worth?" In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. Ed. by Nicoletta Calzolari (Conference Chair) et al. Portorož, Slovenia: European Language Resources Association (ELRA). ISBN: 978-2-9517408-9-1.

Sang, Erik Tjong Kim and Johan Bos (2012). "Predicting the 2011 Dutch senate election results with Twitter." In: *Proceedings of the workshop on semantic analysis in social media*. Association for Computational Linguistics, pp. 53–60.

Saralegi, X. and I. San Vicente (2013). "Elhuyar at TASS 2013." In: *Proceedings of the Workshop on Sentiment Analysis at SEPLN (TASS 2013)*, pp. 143–150.

Saralegi, Xabier and Iñaki San Vicente (2013). "Elhuyar at TASS 2013." In: *XXIX Congreso de la Sociedad Española de Procesamiento de Lenguaje Natural (SEPLN 2013). TASS 2013 - Workshop on Sentiment Analysis*

at SEPLN 2013. Ed. by Alberto Díaz Esteban, Iñaki Alegría Loinaz, and Julio Villena Román. Madrid, Spain, pp. 143–150.

Schmid, Helmut (1994). "Part-of-speech tagging with neural networks." In: *Proceedings of the 15th conference on Computational linguistics-Volume 1*. Association for Computational Linguistics, pp. 172–176.

Scholz, Thomas and Stefan Conrad (2013). "Linguistic sentiment features for newspaper opinion mining." In: *International Conference on Application of Natural Language to Information Systems*. Springer, pp. 272–277.

Severyn, A and A Moschitti (2015). "UNITN: Training Deep Convolutional Neural Network for Twitter Sentiment Classification." In: *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*. Denver, Colorado: Association for Computational Linguistics, pp. 464–469.

Severyn, Aliaksei, Alessandro Moschitti, Olga Uryupina, Barbara Plank, and Katja Filippova (2016). "Multi-lingual opinion mining on Youtube." In: *Information Processing & Management* 52.1, pp. 46–60.

Sidorov, G., S. Miranda-Jiménez, F. Viveros-Jiménez, A. Gelbukh, N. Castro-Sánchez, F. Velásquez, I. Díaz-Rangel, S. Suárez-Guerra, A. Treviño, and J. Gordon (2013). "Empirical study of machine learning based approach for opinion mining in tweets." In: *Proceedings of the 11th Mexican international conference on Advances in Artificial Intelligence - Volume Part I*. MICAI'12. San Luis Potosí, Mexico: Springer-Verlag, pp. 1–14. ISBN: 978-3-642-37806-5. DOI: 10.1007/978-3-642-37807-2_1.

Silió, Elisa (2014). *Tania Sánchez reclama una nueva dirección de IU de Madrid que refleje su triunfo en las primarias*. Ed. by www.elpais.com.

Silva, M. J, P. Carvalho, L. Sarmento, E. de Oliveira, and P. Magalhaes (2009). "The design of OPTIMISM, an opinion mining system for Portuguese politics." In: *New trends in artificial intelligence: Proceedings of EPIA*, pp. 12–15.

Socher, Richard, Brody Huval, Christopher D Manning, and Andrew Y Ng (2012). "Semantic compositionality through recursive matrix-vector spaces." In: *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Association for Computational Linguistics, pp. 1201–1211.

Socher, Richard, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Ng, and Christopher Potts (2013). "Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank." In: *EMNLP 2013. 2013 Conference on Empirical Methods in Natural Language Processing. Proceedings of the Conference*. ACL. Seattle, Washington, USA, pp. 1631–1642.

Søgaard, Anders (2011). "Semisupervised condensed nearest neighbor for part-of-speech tagging." In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Lan-*

guage Technologies: short papers-Volume 2. Association for Computational Linguistics, pp. 48–52.

Solorio, Thamar et al. (2014). "Overview for the first shared task on language identification in code-switched data." In: *Proceedings of The First Workshop on Computational Approaches to Code Switching*. Doha, Qatar: Association for Computational Linguistics, pp. 62–72.

Souza, M. and R. Vieira (2012). "Sentiment analysis on Twitter data for Portuguese language." In: *International Conference on Computational Processing of the Portuguese Language*. Springer, pp. 241–247.

Souza, M., R. Vieira, D. Busetti, R. Chishman, I. M. Alves, and Others (2011). "Construction of a Portuguese opinion lexicon from multiple resources." In: *8th Brazilian Symposium in Information and Human Language Technology*, pp. 59–66.

Spencer, J and G Uchyigit (2012). "Sentimentor: Sentiment Analysis on Twitter Data." In: *The 1st International Workshop on Sentiment Discovery from Affective Data*. Bristol, United Kingdom.

Srivastava, Nitish, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov (2014). "Dropout: A simple way to prevent neural networks from overfitting." In: *The Journal of Machine Learning Research* 15.1, pp. 1929–1958.

Taboada, M, J Brooke, M Tofiloski, K Voll, and M Stede (2011). "Lexicon-based methods for sentiment analysis." In: *Computational Linguistics* 37.2, pp. 267–307. ISSN: 0891-2017. DOI: 10.1162/COLI_a_00049.

Taboada, Maite and Jack Grieve (2004). "Analyzing appraisal automatically." In: *Proceedings of AAAI Spring Symposium on Exploring Attitude and Affect in Text (AAAI Technical Report), Stanford University, CA, pp. 158q161. AAAI Press*.

Tang, Duyu, Furu Wei, Bing Qin, Ting Liu, and Ming Zhou (2014). "Coooolll: A deep learning system for Twitter sentiment classification." In: *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pp. 208–212.

Taulé, M., M. A. Martí, and M. Recasens (2008). "AnCora: Multilevel Annotated Corpora for Catalan and Spanish." In: *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*. Ed. by Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odjik, Stelios Piperidis, and Daniel Tapias. Marrakech, Morocco, pp. 96–101. ISBN: 2-9517408-4-0.

Thelwall, M, K Buckley, and G Paltoglou (2012). "Sentiment strength detection for the social web." In: *J. Am. Soc. Inf. Sci. Technol.* 63.1, pp. 163–173.

Thelwall, Mike, Kevan Buckley, Georgios Paltoglou, Di Cai, and Arvid Kappas (2010). "Sentiment Strength Detection in Short Informal

Text." In: *Journal of the American Society for Information Science and Technology* 61.12, pp. 2544–2558.

Titov, Ivan and James Henderson (2007). "Constituent parsing with incremental sigmoid belief networks." In: *Annual Meeting of the Association for Computational Linguistics*. Vol. 45. 1, p. 632.

Toutanova, K and C D Manning (2000). "Enriching the knowledge sources used in a maximum entropy part-of-speech tagger." In: *Proceedings of the 2000 Joint SIGDAT conference on Empirical methods in natural language processing and very large corpora: held in conjunction with the 38th Annual Meeting of the Association for Computational Linguistics-Volume 13*, pp. 63–70.

Trnavac, Radoslava and Maite Taboada (2012). "The contribution of nonveridical rhetorical relations to evaluation in discourse." In: *Language Sciences* 34.3, pp. 301–318.

Tumasjan, Andranik, Timm O Sprenger, Phillip G Sandner, and Isabell M Welpe (2010). "Predicting Elections in Twitter: What 140 Characters Reveal about Political Sentiment." In: *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media (ICWSM 2010)*. Ed. by William W Cohen and Samuel Gosling. AAAI. Washington, DC, pp. 178–185.

Turney, P D (2002). "Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews." In: *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*. ACL '02. Stroudsburg, PA, USA: ACL, pp. 417–424. DOI: 10.3115/1073083.1073153.

Turney, Peter (2001). "Mining the web for synonyms: PMI-IR versus LSA on TOEFL." In: *European Conference on Machine Learning*.

Vilares, David, Miguel A. Alonso, and Carlos Gómez-Rodríguez (2013). "Clasificación de polaridad en textos con opiniones en español mediante análisis sintáctico de dependencias." In: *Procesamiento de Lenguaje Natural* 50, pp. 13–20.

Vilares, David, Miguel A. Alonso, and Carlos Gómez-Rodríguez (2013). "LyS at TASS 2013: Analysing Spanish tweets by means of dependency parsing, semantic-oriented lexicons and psychometric word-properties." In: *Proc. of the TASS workshop at SEPLN 2013. IV Congreso Español de Informática*, pp. 179–186.

Vilares, David, Miguel A. Alonso, and Carlos Gómez-Rodríguez (2013). "Supervised polarity classification of Spanish tweets based on linguistic knowledge." In: *DocEng'13. Proceedings of the 13th ACM Symposium on Document Engineering*. ACM. Florence, Italy, pp. 169–172.

Vilares, David, Miguel A. Alonso, and Carlos Gómez-Rodríguez (2015a). "A syntactic approach for opinion mining on Spanish reviews." In: *Natural Language Engineering* 21.01, pp. 139–163.

— (2015b). "On the usefulness of lexical and syntactic processing in polarity classification of Twitter messages." In: *Journal of the*

Association for Information Science and Technology 66.9, pp. 1799–1816.

Vilares, David, Miguel A. Alonso, and Carlos Gómez-Rodríguez (2015). "Sentiment Analysis on Monolingual, Multilingual and Code-Switching Twitter Corpora." In: *Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. Lisboa, Portugal: Association for Computational Linguistics, pp. 2–8.

Vilares, David, Miguel A. Alonso, and Carlos Gómez-Rodríguez (2016). "EN-ES-CS: An English-Spanish Code-Switching Twitter Corpus for Multilingual Sentiment Analysis." In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. Portorož, Slovenia: European Language Resources Association (ELRA), pp. 4149–4153.

Vilares, David, Miguel A. Alonso, and Carlos Gómez-Rodríiguez (2013). "Una aproximación supervisada para la minería de opiniones sobre tuits en español en base a conocimiento lingüístico." In: *Procesamiento del lenguaje natural* 51, pp. 127–134. ISSN: 1135-5948.

Vilares, David and Miguel Alonso (2016). "A review on political analysis and social media." In: *Procesamiento del Lenguaje Natural* 56, pp. 13–24.

Vilares, David, Carlos Gómez-Rodríguez, and Miguel A. Alonso (2016). "One model, two languages: training bilingual parsers with harmonized treebanks." In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Berlin, Germany: Association for Computational Linguistics, pp. 425–431.

Vilares, David, Carlos Gómez-Rodríguez, and Miguel A. Alonso (2017a). "Supervised Sentiment Analysis in Multilingual Environments." In: *Information Processing & Management* 53, pp. 595–607.

— (2017b). "Universal, unsupervised (rule-based), uncovered sentiment analysis." In: *Knowledge-Based Systems* 118, pp. 45–55.

Vilares, David, Mike Thelwall, and Miguel A. Alonso (2015). "The megaphone of the people? Spanish SentiStrength for real-time analysis of political tweets." In: *Journal of Information Science* 41.6, pp. 799–813.

Vilares, David, Miguel Hermo, Miguel A. Alonso, Carlos Gómez-Rodríguez, and Yerai Doval (2014a). "LyS: Porting a Twitter Sentiment Analysis Approach from Spanish to English." In: *Proceedings og The 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pp. 411–415.

Vilares, David, Miguel Hermo, Miguel A. Alonso, Carlos Gómez-Rodríguez, and Jesús Vilares (2014b). "LyS at CLEF RepLab 2014: Creating the state of the art in author influence ranking and reputation classification on Twitter." In: *Proceedings of the Fifth International Conference of the CLEF initiative*, pp. 1468–1478.

Vilares, David, Yerai Doval, Miguel A. Alonso, and Carlos Gómez-Rodrıguez (2014c). "Lys at TASS 2014: A prototype for extracting and analysing aspects from Spanish tweets." In: *Proceedings of the TASS workshop at SEPLN*.

Vilares, David, Yerai Doval, Miguel A. Alonso, and Carlos Gómez-Rodríguez (2016). "LyS at SemEval-2016 Task 4: Exploiting Neural Activation Values for Twitter Sentiment Classification and Quantification." In: *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*. San Diego, California: Association for Computational Linguistics, pp. 79–84.

Villena-Román, Julio and Janine García-Morera (2013). "TASS 2013 — Workshop on Sentiment Analysis at SEPLN 2013: An overview." In: *XXIX Congreso de la Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN 2013). TASS 2013 - Workshop on Sentiment Analysis at SEPLN 2013*. Ed. by A Díaz Esteban, I Alegría Loinaz, and J Villena Román. Madrid, Spain, pp. 112–125.

Villena-Román, Julio, Sara Lana-Serrano, Eugenio Martínez-Cámara, and José Carlos González-Cristóbal (2013). "TASS — Workshop on Sentiment Analysis at SEPLN." In: *Procesamiento del Lenguaje Natural* 50, pp. 37–44.

Vinodhini, G and RM Chandrasekaran (2012). "Sentiment analysis and opinion mining: a survey." In: *International Journal of Advanced Research in Computer Science and Software Engineering* 2.6, pp. 282–292.

Volkova, Svitlana, Theresa Wilson, and David Yarowsky (2013). "Exploring Sentiment in Social Media: Bootstrapping Subjectivity Clues from Multilingual Twitter Streams." In: *ACL (2)*, pp. 505–510.

Volokh, Alexander and Günter Neumann (2012). "Task-oriented dependency parsing evaluation methodology." In: *Information Reuse and Integration (IRI), 2012 IEEE 13th International Conference on*. IEEE, pp. 132–137.

Wan, Xiaojun (2009). "Co-training for cross-lingual sentiment classification." In: *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-volume 1*. Association for Computational Linguistics, pp. 235–243.

Wang, Hao, Dogan Can, Abe Kazemzadeh, François Bar, and Shrikanth Narayanan (2012). "A system for real-time twitter sentiment analysis of 2012 us presidential election cycle." In: *Proceedings of the ACL 2012 System Demonstrations*. Association for Computational Linguistics, pp. 115–120.

Wang, Yequan, Minlie Huang, Xiaoyan Zhu, and Li Zhao (2016). "Attention-based LSTM for Aspect-level Sentiment Classification." In: *EMNLP*.

Wiebe, J., T. Wilson, and C. Cardie (2005). "Annotating expressions of opinions and emotions in language." In: *Language resources and evaluation* 39, pp. 165–210.

Wiebe, Janyce M and Rebecca F Bruce (2001). "Probabilistic classifiers for tracking point of view." In: *PROGRESS IN COMMUNICATION SCIENCES*, pp. 125–142.

Wiebe, Janyce M., Rebecca F. Bruce, and Thomas P. O'Hara (1999). "Development and Use of a Gold-standard Data Set for Subjectivity Classifications." In: *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics*. ACL '99. College Park, Maryland: Association for Computational Linguistics, pp. 246–253. ISBN: 1-55860-609-3.

Wilks, Yorick and Janusz Bien (1983). "Beliefs, points of view, and multiple environments." In: *Cognitive Science* 7.2, pp. 95–119.

Wu, Yuanbin, Qi Zhang, Xuanjing Huang, and Lide Wu (2009). "Phrase Dependency Parsing for Opinion Mining." In: *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*. ACL. Singapore, pp. 1533–1541.

Xiao, Min and Yuhong Guo (2012). "Multi-View AdaBoost for Multilingual Subjectivity Analysis." In: *COLING 2012. 24th International Conference on Computational Linguistics. Proceedings of COLING 2012: Technical Papers*. Ed. by Martin Kay and Christian Boitet. Mumbai, India, pp. 2851–2866.

Yan, Gonjun, Wu He, Jiancheng Shen, and Chuanyi Tang (2014). "A bilingual approach for conducting Chinese and English social media sentiment analysis." In: *Computer Networks* 75.B, pp. 491–503.

Yu, Yang, Wenjing Duan, and Qing Cao (2013). "The impact of social and conventional media on firm equity value: A sentiment analysis approach." In: *Decision Support Systems* 55.4, pp. 919–926.

Zeman, Daniel, David Mareček, Martin Popel, Loganathan Ramasamy, Jan Štěpánek, Zdeněk Žabokrtský, and Jan Hajič (2012). "HamleDT: To Parse or Not to Parse?" In: *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*. Istanbul, Turkey: European Language Resources Association (ELRA). ISBN: 978-2-9517408-7-7.

Zhang, Changli, Daniel Zeng, Jiexum Li, Fei-Yue Wang, and Wanli Zuo (2009). "Sentiment Analysis of Chinese Documents: From Sentence to Document Level." In: *Journal of the American Society for Information Science and Technology* 60.12, pp. 2474–2487.

Zhou, Bolei, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva (2014). "Learning deep features for scene recognition using places database." In: *Advances in neural information processing systems*, pp. 487–495.