Universal, Unsupervised (Rule-Based), Uncovered Sentiment Analysis*

David Vilares, Carlos Gómez-Rodríguez and Miguel A. Alonso
Grupo LyS, Departamento de Computación, Universidade da Coruña
Campus de A Coruña s/n, 15071, A Coruña, Spain
{david.vilares,carlos.gomez,miguel.alonso}@udc.es

Abstract

We present a novel unsupervised approach for multilingual sentiment analysis driven by compositional syntax-based rules. On the one hand, we exploit some of the main advantages of unsupervised algorithms: (1) the interpretability of their output, in contrast with most supervised models, which behave as a black box and (2) their robustness across different corpora and domains. On the other hand, by introducing the concept of compositional operations and exploiting syntactic information in the form of universal dependencies, we tackle one of their main drawbacks: their rigidity on data that are structured differently depending on the language concerned. Experiments show an improvement both over existing unsupervised methods, and over state-of-the-art supervised models when evaluating outside their corpus of origin. Experiments also show how the same compositional operations can be shared across languages. The system is available at `http://www.grupolys.org/software/UUUSA/`.

1 Introduction

Sentiment Analysis (SA) is a subfield of natural language processing (NLP) that deals with the automatic comprehension of the opinions shared by users in different media (Pang and Lee, 2008; Cambria, 2016). One of the main challenges addressed by SA focuses on emulating the semantic composition process carried out by humans when understanding the sentiment of an opinion (i.e., if it is favorable, unfavorable or neutral). In the sentence *'He is not very handsome, but he has something that I really like'*, humans have 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'*. Based on this, a human could justify a positive overall sentiment on that sentence.

Our main contribution is the introduction of the first universal and unsupervised (knowledge-based) model for compositional sentiment analysis (SA) driven by syntax-based rules. We introduce 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 et al. (2012) and McDonald et al. (2013): universal annotation criteria that can be used to represent the morphology and syntax of any language in a uniform way. The model 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.

The remainder of this article is structured as follows. §2 reviews related work. §3 introduces the formalism for compositional operations, which is used in §4 to define a set of universal rules that can process relevant linguistic phenomena for SA in any language. §5 presents experimental results of our approach on different corpora and languages. Finally, §6 concludes and discusses directions for future work.

2 Related work

In this section we describe previous work relevant to the topics covered in this article: the issue of multilinguality in SA, semantic composition through machine learning models and semantic composition on knowledge-based systems.

2.1 Multilingual SA

Monolingual sentiment analysis systems have been created for languages belonging to a variety of language families, such as Afro-Asiatic (Aldayel and Azmi, Forthcoming), Indo-European (Vilares et al., 2015a; Vilares et al., 2015b; Ghorbel and Jacot, 2011; Scholz and Conrad, 2013; Neri et al., 2012; Habernal et al., 2014; Medagoda et al., 2013; Medagoda et al., 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 lexicons are scarce for languages other than English, and therefore a great deal of effort has been dedicated to building lexical resources for sentiment analysis (Kim et al., 2009; Hogenboom et al., 2014; Cruz et al., 2014; Volkova et al., 2013; Gao et al., 2013; Chen and Skiena, 2014). A common approach for obtaining a lexicon for a new language consists in translating pre-existent English lexicons (Brooke et al., 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; Wan, 2009; Perea-Ortega et al., 2013; Martínez Cámara et al., 2014). 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; Balahur and Turchi, 2014). To deal with this issue, several methods have been proposed to reduce translation errors, such as applying both directions of translation simultaneously (Hajmohammadi et al., 2014) or enriching the MT system with sentiment patterns (Hiroshi et al., 2004). In the case of supervised systems, self-training and co-training techniques have also been explored to improve performance (Gui et al., 2013; Gui et al., 2014).

Few multilingual systems for SA tasks have been described in the literature. Banea et al. (Banea et al., 2010; Banea et al., 2014) describe 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 (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. (Yan et al., 2014) describe a supervised multilingual system for SA working on previously tokenized Chinese and English texts. Vilares et al. (2015c) present a multilingual SA system trained on a multilingual dataset that is able to outperform monolingual systems on some monolingual datasets and that can work successfully on code-switching texts, i.e., texts that contain terms written in two or more different languages (Vilares et al., 2016a). Some approaches rely on MT to deal with multi-linguality. Balahur et al. (Balahur et al., 2014) build 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. They also point out that the performance of the monolingual Spanish SA system trained on Spanish machine translated data can be improved by adding native Spanish data for training from the Spanish TASS 2013 Twitter dataset (Villena-Román et al., 2014). In contrast, Balahur and Perea-Ortega (Balahur and Perea-Ortega, 2015) inform that performance decreases when machine-translated English data is used to enlarge the TASS 2013 training corpus for Spanish sentiment analysis.

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 et al., 2012). The use of other language-independent indicators, such as character and punctuation repetitions, results in low recall (Cui et al., 2011).

2.2 Composition in machine learning SA systems

A naïve approach to emulating the comprehension of the meaning of multiword phrases for SA consists in using *n*-grams of words, with $n > 1$ (Pang et al., 2002). The approach is limited by the curse of dimensionality, although crawling data from the target domain can help to reduce that problem (Kiritchenko et al., 2014). Joshi and Penstein-Rosé (2009) went one step forward and proposed generalized dependency triplets as features for subjectivity detection, capturing non-local relations. 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 (Pang and Lee, 2005). Kalchbrenner et al. (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 (Rosenthal et al., 2015), although crawling tens of millions of messages was first required to achieve state-of-the-art results. With a different purpose, Poria et al. (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 is 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 options.

2.3 Composition in knowledge-based SA systems

When the said limitations of machine learning models need to be addressed, unsupervised approaches are useful. In this line, Turney (2002) proposed an unsupervised learning algorithm to calculate the semantic orientation (SO) of a word. Taboada et al. (2011) presented a lexical rule-based approach to handle relevant linguistic phenomena such as intensification, negation, *'but'* clauses and *irrealis*. Thelwall et al. (2012) released SentiStrength, a multilingual unsupervised system for micro-text SA that handles negation and intensification, among other web linguistic phenomena. It is limited to snippet-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 et al. (2009) define a set of syntax-based rules for handling negation in English. Vilares et al. (2015a) propose a syntactic SA method, but limited to Spanish reviews and Ancora trees (Taulé et al., 2008). Cambria et al. (2014) release 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 (AI), 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 select syntactical rules for an unsupervised aspect extraction approach, showing the utility of rule-based systems on opinion mining tasks.

In brief, most unsupervised approaches are language-dependent, and those that can manage multilinguality, such as SentiStrength, cannot apply semantic composition.

3 Unsupervised Compositional SA

In contrast with previous work, we propose 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.

3.1 Operations for compositional SA

Let w=$w_1, ..., w_n$ be a sentence, where each word occurrence $w_i \in W$.

Definition 1. *A* **tagged sentence** *is a list of tuples (w_i, t_i) where each w_i is assigned a part-of-speech tag, t_i, indicating its grammatical category (e.g. noun, verb or adjective).*

Definition 2. *A* **dependency tree** *for 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.*

Example 1. *Figure 1 shows a valid dependency tree for our running example.*

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. 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,
- $lm\text{-}branch_T(i, d) = min\{k \in V \mid i \xrightarrow{d} k\}$, 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.*

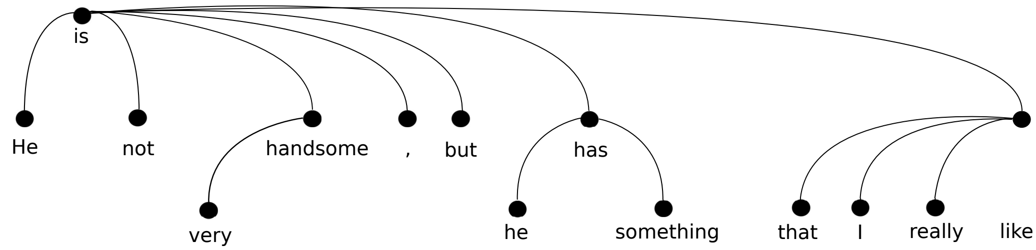

Figure 1: Example of a valid dependency tree for our introductory sentence: *'He is not very handsome, but he has something that I really like'*, following the McDonald et al. (2013) guidelines. For simplicity, we omit the dummy root in the figures.

In practice, our system defines $C(i)$ by means of sets of words, part-of-speech tags and/or dependency types such that the operation will be triggered if w_i, t_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 showed in §5.

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 2):

- $dest$ (*destination node*): The transformation τ is applied directly to the SO of $ancestor_T(i, \delta)$ (see Figure 2.a).
- $lm\text{-}branch^d$ (*branch of* d): The affected nodes are $lm\text{-}branch_T\ (ancestor_T(i,\ \delta), d)$ (see Figure 2.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 2.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 2.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 2.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 2.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 § 3.2), in this paper 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

[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 article, β's for intensifiers are directly taken from existing lexical resources and are not tuned in any way, as explained in §5.

[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.

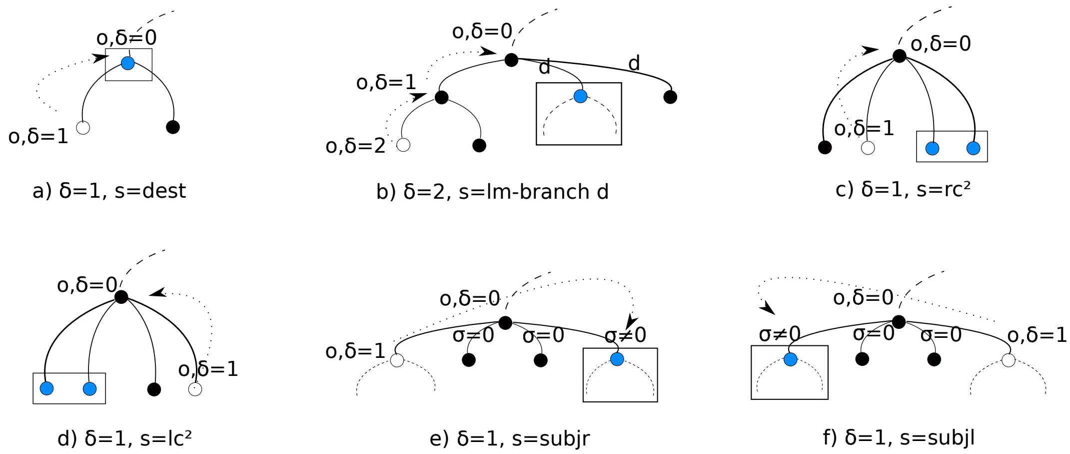

Figure 2: Graphical representation of the proposed set of influence scopes S. ◯ indicates the node that triggers an operation o, □ the nodes to which it is applied (colored in blue).

sentiment lexica which are available for many languages (§ 3.2), so it does not prevent our system from working on a wide set of languages without any adaptation, apart from modifying the subjective lexicon.

3.2 An algorithm for unsupervised SA

Algorithm 1 Compute SO of a node

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

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 $children_T(i)$. 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.

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

- `<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).

3.3 NLP tools for universal unsupervised SA

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

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.). For part-of-speech tagging we rely on the free distribution of the Toutanova and Manning (2000) tagger. Dependency parsers are built using MaltParser (Nivre et al., 2007) and MaltOptimizer (Ballesteros and Nivre, 2012). We trained a set of taggers and parsers for different languages using the universal tag and dependency sets (Petrov et al., 2012; McDonald et al., 2013). In particular, we are relying on the monolingual models using universal part-of-speech tags presented by Vilares et al. (2016b).

With respect to multilingual lexical resources, 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, 2006), where each synset from WordNet is assigned a 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'*).

4 Defining 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 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, the same rules can be competitive across different languages.

4.1 Intensification

Intensification amplifies or diminishes the sentiment of a word or phrase. Simple cases of this phenomenon can be *'I have **huge** problems'* or *'This is **a bit** dissapointing'*. Traditional lexicon-based methods handle most of these cases with simple heuristics (e.g. amplifying or diminishing the sentiment of the word following an intensifier). 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).

Universal compositional operations overcome this problem 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 3 shows graphically how universal dependencies represent the cases discussed above these lines. Formally, the operation for these forms of intensification is: $(weighting_{\beta}, w \in \text{intensifiers} \wedge t \in \{\text{ADV,ADJ}\} \wedge d \in \{\text{advmod,amod,nmod}\}, 1, 3, dest \cup lm\text{-}branch^{\text{acomp}})$, with the value of β depending on the strength of the intensifier as given by the sentiment lexicon.

4.1.1 *'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 authors such as Brooke et al.

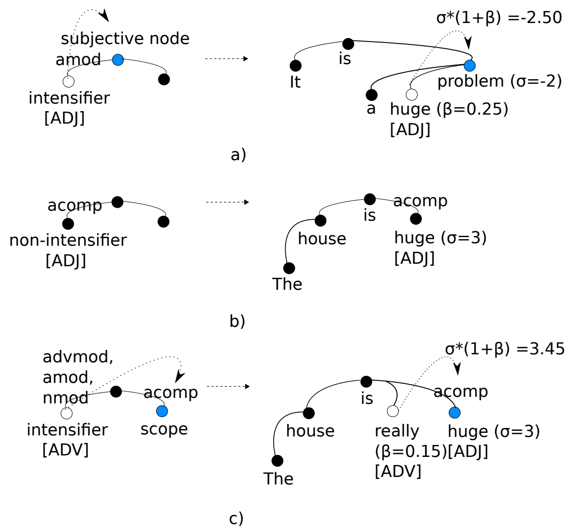

Figure 3: Skeleton for intensification compositional operations (2.a, 2.c) and one case without intensification (2.b), together with examples annotated with universal dependencies. Semantic orientation values are for instructional purposes only. In 2.a, *'huge'* is a term considered in a list of intensifiers, 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+β). 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.

(2009) or Vilares et al. (2015a). 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 4 shows our proposed composition operation for this clause, formally: $(weighting_\beta, w \in \{\text{but}\} \wedge t \in \{\text{CONJ}\} \wedge d \in \{\text{cc}\}, 1, 1, subjl)$ 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.

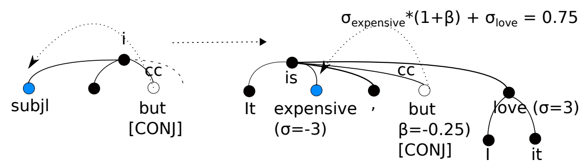

Figure 4: Skeleton for *'but'* compositional operation illustrated with one example according to universal dependencies. The term *'but'* matches the word form, tag and dependency types required to act as a sentence intensifier, so the compositional operation is queued to be applied $\delta = 1$ levels 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+β) (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.

4.2 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. 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 in Figure 5. Formally, the proposed compositional operation to tackle most forms of negation under universal guidelines is: $(shift_\alpha, w \in \text{negations} \wedge t \in U \wedge d \in \{\text{neg}\}, 1, 2, dest \cup lm\text{-}branch^{\text{attr}} \cup lm\text{-}branch^{\text{acomp}} \cup subjr)$, where

U represents the universal tag set. 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.

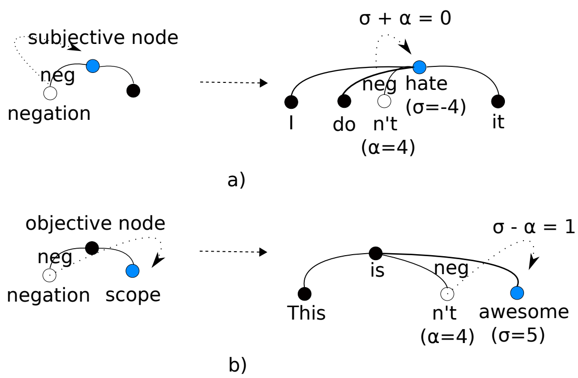

Figure 5: Skeleton for negation compositional operations illustrated together with one example. In 5.a, the term *'n't'* matches the form word of a negator and its dependency type is *neg*, queuing a negation compositional operation to be applied $\delta = 1$ levels upward (i.e., at the *'hate'* node). 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 $shift_\alpha(\sigma)$ 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., $lm\text{-}branch^{\text{attr}}$).

4.3 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: $(weighting_\beta, w \in \{\text{if}\} \wedge t \in U \wedge d \in \{\text{mark}\}, 2, 3, dest \cup subjr)$. Its graphical representation would be very similar to intensification (see Figures 2 a) and e)).

4.4 Discussion

Figure 6 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.[3] Additionally, Table 1 illustrates the internal state of the algorithm and the SO updates made at each step for our running example.

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	[]	$[\text{N}_{not(1,2)}]$	0	0
3	very_4	[]	$[\text{I}_{very(1,3)}]$	0	0
4	handsome_5	$[\text{I}_{very(0,3)}]$	[]	4	5
5	$,_6$	[]	[]	0	0
5	but_7	[]	$[\text{I}_{but(1,1)}]$	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}	[]	$[\text{I}_{really(1,3)}]$	0	0
12	like_{13}	$[\text{I}_{really(1,3)}]$	[]	1	1.15
13	is_2	$[\text{N}_{not(0,2)}, \text{I}_{but(0,1)}]$	[]	0	1.90

Table 1: Internal state and SO updates made by the proposed algorithm for the running example. Each row corresponds to a step in which a node (Word_{index}) is visited in the postorder traversal. Columns $\text{A}_{word(\delta,\pi)}$ and $\text{Q}_{word(\delta,\pi)}$ 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) or Vilares et al. (2015a) is that they present predefined sets of linguistic patterns for language-specific SA, whereas our approach is

[3]The system released together with this paper shows an equivalent ASCII text representation that can be obtained on the command line. It is also possible to check how the system works at https://miopia.grupolys.org/demo/

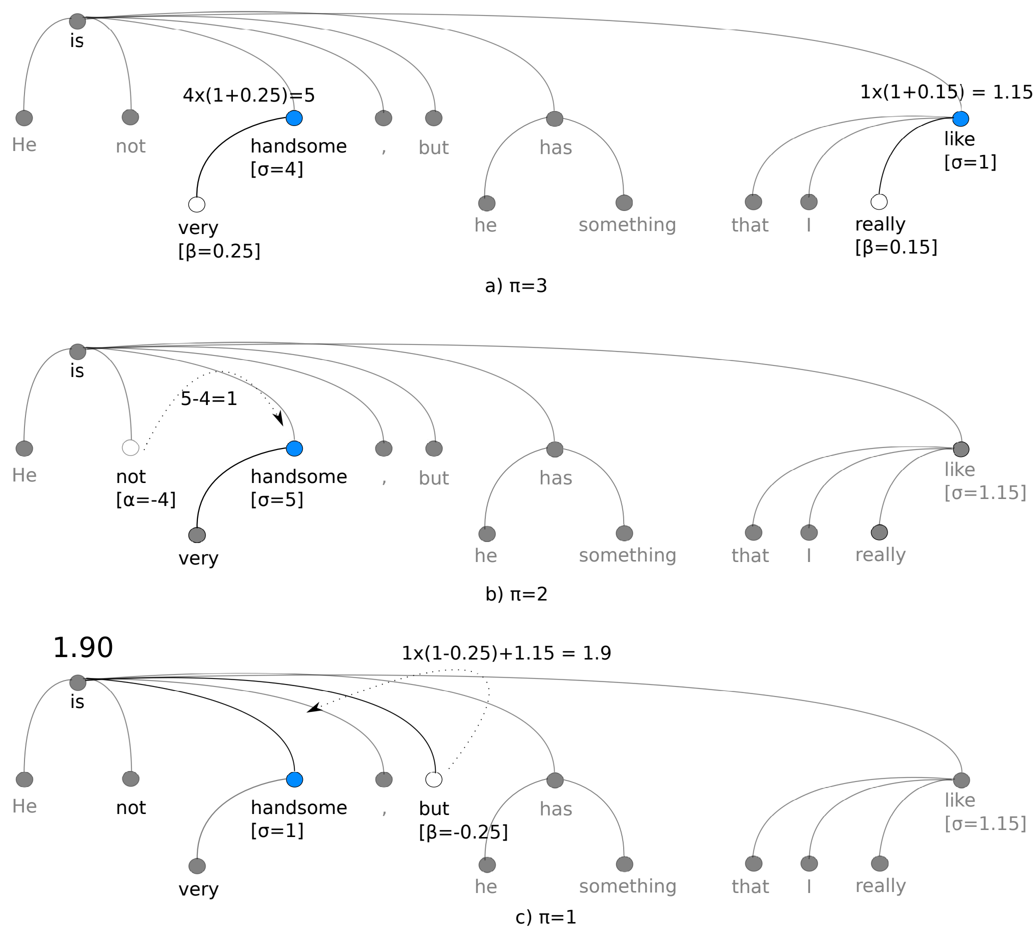

Figure 6: Analysis of a sentence applying universal unsupervised prediction. For the sake of clarity, the real post-order traversal is not illustrated. Instead we show an (in this case) equivalent computation by applying all operations with a given priority, π, at the same time, irrespective of the node. 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.

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).

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).[4]

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[5] 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 (Vilares et al., 2016b; Ammar et al., 2016; Guo et al., 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 et al., 2016b).

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.

[4]There is a difference between the number of compositional operations that are defined in the system (one for each phenomenon considered: intensification, *'but'* clauses, negation and irrealis), and the number of compositional operation instances created at runtime given such definitions. The latter depends on the words, tags and dependency types that match each operation's predicate C. While the matching tags and dependency types are fixed and common to all languages, the number of words that can match C depends on the lexicon, so the number of operation instances varies across languages depending on the use of SO-CAL and SentiStrength as lexical resources (e.g. English is the language that generates more instances, with 1411 compositional operations, due to having the largest intensifier lexicon among the languages and resources considered).

[5]Labeled Attachment Score (LAS): The percentage of dependencies where both the head and the dependency type have been assigned correctly. The English model used has a LAS of 89.36%.

5 Experimental results

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 by Vilares et al. (2015a), an unsupervised syntax-based approach available for Spanish but, in contrast to ours, heavily language-dependent.

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.

5.1 Resources

We selected the following standard English corpora for evaluation:

- Taboada and Grieve (2004) corpus: 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 (Pang and Lee, 2004): A corpus of 2 000 long movie reviews (1 000 positive, 1 000 negative).
- Pang and Lee 2005 corpus (Pang and Lee, 2005): 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 et al. (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 et al. (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.[6] The list of emoticons from Sentistrength 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.[7] The content of these dictionaries and their parameters are not modified or tuned.

[6]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).

5.2 Comparison to unsupervised approaches

Table 2 compares the performance of our model with respect to SentiStrength[8] 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 3 compares these three unsupervised systems on the Pang and Lee 2004 corpus (Pang and Lee, 2004), 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**	**Our system**
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 2: Accuracy (%) on the Taboada and Grieve (2004) corpus. We only provide one row for SentiStrength since we are using the standard configuration for English (which already includes negation and intensification functionalities).

Rules	**SentiStrength**	**SO-CAL**	**Our system**
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 3: Accuracy (%) on Pang and Lee 2004 test set (Pang and Lee, 2004).

Table 4 compares the performance of our universal approach on a different language (Spanish) with respect to: Spanish SentiStrength (Vilares et al., 2015d), the Spanish SO-CAL (Brooke et al.,

[7]Out-of-vocabulary words are not given a special treatment at the moment.

[8]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	Our system	Vilares et al. (2015a)
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 4: Accuracy (%) on the Spanish Brooke et al. (2009) test set.

2009) and a syntactic language-dependent system inspired on the latter (Vilares et al., 2015a). 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 §3.1). Our universal system again outperforms SentiStrength and SO-CAL in its Spanish version. The system also obtains results very similar to the ones reported by Vilares et al. (2015a), 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 article 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 article.

We use the German SentiStrength system (Momtazi, 2012) for comparison. The use of the German SentiStrength dictionary, as mentioned in Section 5.1, 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.

5.3 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 applications, 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 5 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[9] (*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.[10]

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.

6 Conclusions and future work

In this article, we have described, implemented and evaluated a novel model for universal and un-

[9]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.

[10]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)	Our system
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 5: Accuracy (%) on different corpora for Socher et al. (2013) and our system. 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)

supervised sentiment analysis driven by a set of syntactic rules 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.

As future work, we plan to design algorithms for the automatic extraction of compositional operations that capture the semantic relations between tree nodes. We would also like to collect corpora to extend our evaluation to more languages, since collections that are directly available on the web are scarcer than expected. We plan to pay special attention to Sino-Tibetan and Afro-Asiatic languages. With respect to ungrammatical texts, we plan to integrate Tweebo parser (Kong et al., 2014) into our system. Although it does not follow universal guidelines, it will allow us to define compositional operations specifically intended for English tweets and their particular structure. Additionally, the concept of compositional operations is not limited to generic SA and could be adapted for other tasks such as universal aspect extraction. Finally, we plan to adapt the Poria et al. (2014) sentic patterns as compositional operations, so they can be handled universally.

Acknowledgments

This research is supported by the Ministerio de Economía y Competitividad (FFI2014-51978-C2). David Vilares is funded by the Ministerio de Educación, Cultura y Deporte (FPU13/01180). Carlos Gómez-Rodríguez is funded by an Oportunius program grant (Xunta de Galicia). We thank Roman Klinger for his help in obtaining the German data.

References

Haifa K. Aldayel and Aqil M. Azmi. Forthcoming. Arabic tweets sentiment analysis — a hybrid scheme. *Journal of Information Science*. DOI: 10.1177/0165551515610513.

W. Ammar, G. Mulcaire, M. Ballesteros, C. Dyer, and N. A. Smith. 2016. One Parser, Many Languages. *arXiv preprint arXiv:1602.01595*.

Yui Arakawa, 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. *Journal of the Association for Information Science and Technology*, 65(7):1416–1423, July.

A. Aue and M. Gamon. 2005. Customizing sentiment classifiers to new domains: A case study. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP)*.

Alexandra Balahur and José M. Perea-Ortega. 2015. Sentiment analysis system adaptation for multilingual processing: The case of tweets. *Information Processing and Management*, 51(4):547–556, July.

Alexandra Balahur and Marco Turchi. 2012a. Comparative experiments for multilingual sentiment analysis using machine translation. In Mohamed Medhat Gaber, Mihaela Cocea, Stephan Weibelzahl, Ernestina Menasalvas, and Cyril Labbe, editors, *SDAD 2012, The 1st International Workshop on Sentiment Discovery from Affective Data*, pages 75–86, Bristol, UK.

Alexandra Balahur and Marco Turchi. 2012b. Multilingual sentiment analysis using machine translation? In *WASSA 2012, 3rdWorkshop on Computational Approaches to Subjectivity and Sentiment Analysis, Proceedings of the Workshop*, pages 52–60, Jeju, Republic of Korea.

Alexandra Balahur and Marco Turchi. 2014. Comparative experiments using supervised learning and machine translation for multilingual sentiment analysis. *Computer Speech and Language*, 28(1):56–75, January.

Alexandra Balahur, 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 Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, May. European Language Resources Association (ELRA).

M Ballesteros 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*, pages 58–62. Association for Computational Linguistics.

Carmen Banea, Rada Mihalcea, and Janyce Wiebe. 2010. Multilingual subjectivity: Are more languages better? In Chu-Ren Huang and Dan Jurafsky, editors, *COLING 2010. 23rd International Conference on Computational Linguistics. Proceedings of the Conference*, volume 2, pages 28–36, Beijing, August. Tsinghua University Press.

Carmen Banea, Rada Mihalceaa, and Janyce Wiebe. 2014. Sense-level subjectivity in a multilingual setting. *Computer Speech & Language*, 28(1):7–19, January.

Erik Boiy and Marie-Francine Moens. 2009. A machine learning approach to sentiment analysis in multilingual web texts. *Information Retrieval*, 12(5):526–558, October.

J. Brooke, M. Tofiloski, and M. Taboada. 2009. Cross-Linguistic Sentiment Analysis: From English to Spanish. In *Proceedings of RANLP 2009, Recent Advances in Natural Language Processing*, pages 50–54, Bovorets, Bulgaria, September.

Erik Cambria, Amir Hussain, Catherine Havasi, and Chris Eckl. 2009. Affectivespace: blending common sense and affective knowledge to perform emotive reasoning. *WOMSA at CAEPIA, Seville*, pages 32–41.

Erik Cambria, 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*, pages 1515–1521.

Erik Cambria. 2016. Affective computing and sentiment analysis. *IEEE Intelligent Systems*, 31(2):102–107.

Y. Chen and S. 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*, pages 383–389, Baltimore, June. ACL.

Boxing Chen 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*, pages 607–615, Baltimore, June. ACL.

Alex Cheng and Oles Zhulyn. 2012. A system for multilingual sentiment learning on large data sets. In Martin Kay and Christian Boitet, editors, *COLING 2012. 24th International Conference on Computational Linguistics. Proceedings of COLING 2012: Technical Papers*, pages 577–592, Mumbai, India, December.

Md. Faisal Mahbub Chowdhury, 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)*, pages 466–470, Atlanta, Georgia, June. ACL.

Fermín Cruz, José A. Troyano, Beatriz Pontes, and F. Javier Ortega. 2014. Building layered, multilingual sentiment lexicons at synset and lemma levels. *Expert Systems with Applications*, 41(13):5984–5994, October.

Anqi Cui, Min Zhang, Yiqun Liu, and Shaoping Ma. 2011. Emotion tokens: Bridging the gap among multilingual Twitter sentiment analysis. In Mohamed Vall Mohamed Salem, Khaled Shaalan, Farhad Oroumchian, Azadeh Shakery, and Halim Khelalfa, editors, *Information Retrieval Technology. 7th Asia Information Retrieval Societies Conference, AIRS 2011, Dubai, United Arab Emirates, December 18-20, 2011. Proceedings*, volume 7097 of *Lecture Notes in Computer Science*, pages 238–249. Springer, Berlin and Heidelberg.

Alex Davies and Zoubin Ghahramani. 2011. Language-independent bayesian sentiment mining of Twitter. In *The 5th SNA-KDD Workshop'11 (SNA-KDD'11)*, San Diego, CA, August. ACM.

M. De Marneffe and C. D. Manning. 2008. The Stanford typed dependencies representation. In *Coling 2008: Proceedings of the workshop on Cross-Framework and Cross-Domain Parser Evaluation*, pages 1–8. Association for Computational Linguistics.

Erkin Demirtas 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 (WISDOM 2013)*, page Article No. 9, Chicago, IL, USA, Aug. ACM.

Andrea Esuli 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)*, volume 6, pages 417–422, Genoa, Italy.

Dehong Gao, 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*, Bellevue, Washington, USA, July. AAAI.

Hatem Ghorbel and David Jacot. 2011. Sentiment analysis of French movie reviews. In Vincenzo Pallotta, Alessandro Soro, and Eloisa Vargiu, editors, *Advances in Distributed Agent-Based Retrieval Tools*, volume 361 of *Studies in Computational Intelligence*, pages 97–108. Springer, Berlin and Heidelberg.

K. Gimpel, 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 *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pages 42–47. Association for Computational Linguistics.

Lin Gui, 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 Guodong Zhou, Juanzi Li, Dongyan Zhao, and Yansong Feng, editors, *Natural Language Processing and Chinese Computing*, volume 400 of *Communications in Computer and Information Science*, pages 93–104, Heidelberg, NewYork, Dordrecht and London. Springer.

Lin Gui, 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*, pages 860–865, Baltimore, June. ACL.

Jiang Guo, Wanxiang Che, Haifeng Wang, and Ting Liu. 2016. Exploiting multi-typed treebanks for parsing with deep multi-task learning. *arXiv preprint arXiv:1606.01161*.

Ivan Habernal, Tomás Ptáček, and Josef Steinberg. 2014. Supervised sentiment analysis in Czech social media. *Information Processing and Management*, 50(5):693–707, September.

Mohammad Sadegh Hajmohammadi, Roliana Ibrahim, and Ali Selamat. 2014. Bi-view semi-supervised active learning for cross-lingual sentiment classification. *Information Processing and Management*, 50(5):718–732, September.

Kanayama Hiroshi, 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, August.

Alexander Hogenboom, Bas Heerschop, Flavius Frasincar, Uzay Kaymak, and Franciska de Jong. 2014. Multi-lingual support for lexicon-based sentiment analysis guided by semantics. *Decision Support Systems*, 62:43–53, June.

Piyatida Inrak and Sukree Sinthupinyo. 2010. Applying latent semantic analysis to classify emotions in Thai text. In *2nd International Conference on Computer Engineering and Technology (ICCET)*, volume 6, pages 450–454, Chengdu, April. IEEE.

L. Jia, C. Yu, and W. 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*, pages 1827–1830, Hong Kong, November. ACM, ACM Press.

M. Joshi and C. Penstein-Rosé. 2009. Generalizing dependency features for opinion mining. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, ACLShort '09, pages 313–316, Stroudsburg, PA, USA. Association for Computational Linguistics.

N. Kalchbrenner, E. Grefenstette, and P. 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*, pages 655–665, Baltimore, Maryland, USA. ACL.

Jungi Kim, 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 Wenjie Li and Diego Mollá-Aliod, editors, *Computer Processing of Oriental Languages. Language Technology for the Knowledge-based Economy*, volume 5459 of *Lecture Notes in Computer Science*, pages 112–121. Springer, Berlin and Heidelberg.

S. Kiritchenko, X. Zhu, and S. M. Mohammad. 2014. Sentiment Analysis of Short Informal Texts. *Journal of Artificial Intelligence Research*, 50(1):723–762, may.

Roman Klinger and Philipp Cimiano. 2014. The USAGE review corpus for fine-grained, multi-lingual opinion analysis. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, May. European Language Resources Association (ELRA).

Lingpeng Kong, Nathan Schneider, Swabha Swayamdipta, Archna Bhatia, Chris Dyer, and Noah A Smith. 2014. A Dependency Parser for Tweets. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1001–1012, Doha, Qatar, oct. ACL.

Hugo Liu and Push Singh. 2004. Conceptneta practical commonsense reasoning tool-kit. *BT technology journal*, 22(4):211–226.

Qian Liu, Zhiqiang Gao, Bing Liu, and Yuanlin Zhang. 2016. Automated rule selection for opinion target extraction. *Konwledge-Based Systems*, 104:74–88, July.

Eugenio Martínez Cámara, M. Teresa Martín Valdivia, M. Dolores Molina-González, and José M. Perea-Ortega. 2014. Integrating Spanish lexical resources by meta-classifiers for polarity classification. *Journal of Information Science*, 40(4):538–554, August.

R. McDonald, J. Nivre, Y. Quirmbach-brundage, Y. Goldberg, D Das, K. Ganchev, K. Hall, S. Petrov, H. Zhang, O. Täckström, C. Bedini, N. Castelló, and J. Lee. 2013. Universal Dependency Annotation for Multilingual Parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 92–97. Association for Computational Linguistics.

Nishantha Medagoda, Subana Shanmuganathan, and Jacqueline Whalley. 2013. A comparative analysis of opinion mining and sentiment classification in non-English languages. In *Proceedings of ICTER 2013, International Conference on Advances in ICT for Emerging Regions*, Colombo, Sri Lanka, December. IEEE.

Saeedeh Momtazi. 2012. Fine-grained german sentiment analysis on social media. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Uur Doan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 1215–1220, Istanbul, Turkey. European Language Resources Association (ELRA).

Sascha Narr, 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, September.

Federico Neri, Carlo Aliprandi, Federico Capeci, Montserrat Cuadros, and Tomas By. 2012. Sentiment analysis on social media. In *Proceedings of the 2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 951–958. IEEE, August.

J. Nivre, J. Hall, J. Nilsson, A. Chanev, G. Eryigit, S. Kübler, S. Marinov, and E. Marsi. 2007. MaltParser: A language-independent system for data-driven dependency parsing. *Natural Language Engineering*, 13(2):95–135.

B. Pang and L. 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*, pages 271–278. Association for Computational Linguistics.

B. Pang 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*, pages 115–124. Association for Computational Linguistics.

B Pang and L Lee. 2008. *Opinion Mining and Sentiment Analysis*. now Publishers Inc., Hanover, MA, USA.

B Pang, L Lee, and S Vaithyanathan. 2002. Thumbs up? Sentiment classification using machine learning techniques. In *Proceedings of EMNLP*, pages 79–86.

José M. Perea-Ortega, 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. *Journal of the American Society for Information Science and Technology*, 64(9):1864–1877, September.

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2012. A universal part-of-speech tagset. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Uur Doan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may. European Language Resources Association (ELRA).

Soujanya Poria, Erik Cambria, Grégoire Winterstein, and Guang-Bin Huang. 2014. Sentic patterns: Dependency-based rules for concept-level sentiment analysis. *Knowledge-Based Systems*, 69:45–63, October.

Soujanya Poria, Erik Cambria, and Alexander Gelbukh. 2016. Aspect extraction for opinion mining with a deep convolutional neural network. *Knowledge-Based Systems*, 108:42–49.

S. Rosenthal, P. Nakov, S. Kiritchenko, S. M Mohammad, A. Ritter, and V. Stoyanov. 2015. Semeval-2015 task 10: Sentiment analysis in Twitter. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*.

Thomas Scholz and Stefan Conrad. 2013. Linguistic sentiment features for newspaper opinion mining. In *Natural Language Processing and Information Systems. 18th International Conference on Applications of Natural Language to Information Systems, NLDB 2013, Salford, UK, June 19-21, 2013. Proceedings*, volume 7934 of *Lecture Notes in Computer Science*, pages 272–277. Springer, Berlin and Heidelberg.

A Severyn 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)*, pages 464–469, Denver, Colorado. Association for Computational Linguistics.

Aliaksei Severyn, Alessandro Moschitti, Olga Uryupina, Barbara Plank, and Katja Filippova. 2016. Multi-lingual opinion mining on TouTube. *Information Processing and Management*, 52(1):46–60, January.

R. Socher, B. Huval, C. D. Manning, and A. 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*, pages 1201–1211. Association for Computational Linguistics.

R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Ng, and C. 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*, pages 1631–1642, Seattle, Washington, USA, oct. ACL.

M. Taboada and J. Grieve. 2004. Analyzing appraisal automatically. In *Proceedings of AAAI Spring Symposium on Exploring Attitude and Affect in Text (AAAI Technical Report SS0407), Stanford University, CA*, pages 158–161. AAAI Press.

M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede. 2011. Lexicon-based methods for sentiment analysis. *Computational Linguistics*, 37(2):267–307.

M. Taulé, M. A. Martí, and M. Recasens. 2008. AnCora: Multilevel Annotated Corpora for Catalan and Spanish. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odjik, Stelios Piperidis, and Daniel Tapias, editors, *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, pages 96–101, Marrakech, Morocco.

M. Thelwall, K. Buckley, and G. Paltoglou. 2012. Sentiment strength detection for the social web. *J. Am. Soc. Inf. Sci. Technol.*, 63(1):163–173.

K. Toutanova 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*, pages 63–70.

P. D. Turney. 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, pages 417–424, Stroudsburg, PA, USA. ACL.

D. Vilares, M. A. Alonso, and C. Gómez-Rodríguez. 2015a. A syntactic approach for opinion mining on Spanish reviews. *Natural Language Engineering*, 21(1):139–163.

David Vilares, Miguel A. Alonso, and Carlos Gómez-Rodríguez. 2015b. On the usefulness of lexical and syntactic processing in polarity classification of Twitter messages. *Journal of the Association for Information Science and Technology*, 66(9):1799–1816, September.

David Vilares, Miguel A Alonso, and Carlos Gómez-Rodríguez. 2015c. 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*, pages 2–8, Lisboa, Portugal, sep. Association for Computational Linguistics.

David Vilares, Mike Thelwall, and Miguel A Alonso. 2015d. The megaphone of the people? Spanish SentiStrength for real-time analysis of political tweets. *Journal of Information Science*, 41(6):799–813.

David Vilares, Miguel A. Alonso, and Carlos Gómez-Rodríguez. 2016a. En-es-cs: An english-spanish code-switching twitter corpus for multilingual sentiment analysis. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 4149–4153, Portorož, Slovenia, May. ELRA.

David Vilares, Carlos Gómez-Rodríguez, and Miguel A. Alonso. 2016b. 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), pages 425–431, Berlin, Germany, August. Association for Computational Linguistics.

Julio Villena-Román, Janine García-Morera, Cristina Moreno-García, Sara Lana-Serrano, and José Carlos González-Cristóba. 2014. TASS 2013 — a second step in reputation analysis in Spanish. *Procesamiento del Lenguaje Natural*, 52:37–44, March.

G. Vinodhini and RM. Chandrasekaran. 2012. Sentiment analysis and opinion mining: A survey. *International Journal of Advanced Research in Computer Science and Software Engineering*, 2(6):282–292, June.

Svitlana Volkova, Theresa Wilson, and David Yarowsky. 2013. Exploring sentiment in social media: Bootstrapping subjectivity clues from multilingual Twitter streams. In *ACL 2013. 51st Annual Meeting of the Association for Computational Linguistics Proceedings of the Conference. Volume 2: Short Papers*, pages 505–510, Sofia, Bulgaria, August. ACL.

Xiaojun Wan. 2009. Co-training for cross-lingual sentiment classification. In *ACL-IJCNLP 2009. Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and 4th International Joint Conference on Natural Language Processing of the AFNLP. Proceedings of the Conference*, pages 235–243, Suntec, Singapore, August. ACL, World Scientific Publishing Co Pte Ltd.

Min Xiao and Yuhong Guo. 2012. Multi-view AdaBoost for multilingual subjectivity analysis. In Martin Kay and Christian Boitet, editors, *COLING 2012. 24th International Conference on Computational Linguistics. Proceedings of COLING 2012: Technical Papers*, pages 2851–2866, Mumbai, India, December.

Gonjun Yan, Wu He, Jiancheng Shen, and Chuanyi Tang. 2014. A bilingual approach for conducting Chinese and English social media sentiment analysis. *Computer Networks*, 75(B):491–503, December.

Changli Zhang, Daniel Zeng, Jiexum Li, Fei-Yue Wang, and Wanli Zuo. 2009. Sentiment analysis of Chinese documents: From sentence to document level. *Journal of the American Society for Information Science and Technology*, 60(12):2474–2487, December.