Supervised Sentiment Analysis in Multilingual Environments

David Vilares, Miguel A. Alonso and Carlos Gómez-Rodríguez

Grupo LyS, Departamento de Computación, Universidade da Coruña
Campus de A Coruña s/n, 15071, A Coruña, Spain

Abstract

This article tackles the problem of performing multilingual polarity classification on Twitter, comparing 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 techniques were evaluated on monolingual, synthetic multilingual and code-switching corpora of English and Spanish tweets. In the latter case we introduce the first code-switching Twitter corpus with sentiment labels. The samples are labelled according to two well-known criteria used for this purpose: the *SentiStrength scale* and a *trinary scale* (*positive, neutral* and *negative* categories). The experimental results show the robustness of the multilingual approach (1) and also that it outperforms the monolingual models on some monolingual datasets.

Keywords: Sentiment Analysis, Multilingual, Code-Switching.

*Corresponding author: David Vilares

Email address: {david.vilares, miguel.alonso, carlos.gomez}@udc.es (David Vilares, Miguel A. Alonso and Carlos Gómez-Rodríguez)

1. Introduction

Automatically understanding all the information shared on the Web and transforming it into knowledge is one of the main challenges in the age of Big Data. In terms of natural language processing (NLP), this usually involves comprehending different human languages such as English, Spanish or Arabic, which are implicitly related with relevant human aspects such as cultures, countries or even religions. A very simple example of these real differences can be illustrated by the concept *dragon*, which has a positive perception in Chinese, but not necessarily in other languages such as English or Spanish.

In this context, Twitter has become one of the most useful social networks for trending analysis, given the amount of data and its popularity in different countries (Cambria et al., 2013a,b). Some of these trends are global (e.g. the Oscars, Superbowl, Rihanna or the recent Volkswagen scandal) and so their trending topics are also global (e.g. '#oscars2016', '#superbowl2016', ...). However, the public perception of these trends often changes from one culture to another and the task becomes even harder when tweets are written in different languages. This is a challenge for global companies and organizations that need to make specific business and marketing decisions depending on their target population. However, if their monitoring processes are focused on a single language (usually English) the knowledge that they acquire might be incomplete, or even worse, inaccurate. There are even more difficult and unexplored multilingual variants, such as code-switching texts (i.e. texts that contain terms in two or more different languages). Colloquial creole languages such as *Spanglish* (a mix of Spanish and American English) or *Singlish* (English-based creole from Singapore) or even official languages such as the *Haitian creole* (which merges Portuguese, Spanish, Taíno, and West African languages), are some of the best-known situations.

As a result, there is a need to provide effective support for analyzing user-generated content that lacks structure and is created in different languages (Dang et al., 2014). In this context, *sentiment analysis* (SA) techniques have

been successfully applied to this social network in order to monitor a wide variety of issues ranging from the perception of the public with respect to popular events (Thelwall et al., 2011) to political analysis, determining the political opinion of users (Cotelo et al., 2016) or showing whether the sentiment expressed in messages is positive, negative or neutral (Vilares et al., 2015d). However, most of the existing research on sentiment analysis is either monolingual or cross-lingual: models intended for purely multilingual or code-switching messages are scarce. This article fills this gap, describing a novel method for multilingual polarity classification that relies on fusing existing monolingual corpora, instead of applying MT techniques or language-specific pipelines.

This article has the following research objectives:

1. To build the first code-switching corpus from Twitter for sentiment analysis. Each tweet collected in such a corpus will contain words written in at least two different languages.
2. To design a multilingual sentiment analysis system able to determine the sentiment present in texts written in different languages. To do this, we apply soft-data fusion (Khaleghi et al., 2013) at the core level of the information fusion process applied to SA (Level 2 - Situation Refinement), as illustrated by Balazs and Velásquez (2016). In particular, existing monolingual corpora are fused to create such multilingual system.
3. To evaluate the performance of the multilingual sentiment analysis system on standard corpora and on the novel code-switching corpus, comparing its performance with respect to the combination of a language detection system and monolingual sentiment analysis systems.

For these purposes, we will consider English (*en*) and Spanish (*es*) as working languages throughout this article. Thus, the aim of the article is to show how current supervised approaches can address situations where monolingual, multilingual and code-switching texts appear.

The remainder of the paper is organised as follows: Section 2 discusses the state of the art regarding opinion mining on texts in diverse languages, including

monolingual, cross-lingual and multilingual approaches. Section 3 describes the process and result of building the code-switching corpus. Section 4 introduces the main ideas and features of the proposed models. Section 5 defines the experimental framework and outlines the corpora used for evaluation, including both standard collections and the novel code-switching corpus. Section 6 presents the results obtained by the models on these corpora, which are discussed in Section 7. Finally, Section 8 draws our conclusions and outlines for future research.

2. Related Work

We start by considering the issues we must face when mining opinions from non-English texts. We then focus on work applying a given opinion mining technique to corpora in different languages. Next, we review work on cross-language opinion mining and finally we consider work on multilingual subjectivity detection and polarity classification.

2.1. Mining opinions from non-English texts

There is recent work on the definition of language-specific methods for opinion mining in a wide variety of languages, including, among others, Arabic (Aldayel and Azmi, 2015), Chinese (Vinodhini and Chandrasekaran, 2012; Zhang et al., 2009), Czech (Habernal et al., 2014), French (Ghorbel and Jacot, 2011), German (Scholz and Conrad, 2013), Hindi (Medagoda et al., 2013), Italian (Neri et al., 2012), Japanese (Arakawa et al., 2014), Russian (Medagoda et al., 2013), Spanish (Vilares et al., 2015c) and Thai (Inrak and Sinthupinyo, 2010). One of the problems we face when dealing with languages other than English is that many English language sentiment dictionaries are freely available, but such vocabulary lists are scarce for other languages. A current line of work is the automatic or semi-automatic generation of large non-English sentiment vocabularies (Steinberger, 2012). In this line, Kim et al. (2009) propose to create a sentiment lexicon for Korean using two sentiment lexicons for English, a bilingual dictionary and a link analysis algorithm. Hogenboom et al. (2014) propose

to project sentiment scores from the English SentiWordNet (Baccianella et al., 2010) to Dutch. In the same line, Cruz et al. (2014) use MCR (Gonzalez-Agirre et al., 2012) and EuroWordNet (Vossen, 1998) to transfer sentiment from the English SentiWordNet to the Spanish, Catalan, Galician and Basque WordNets.

Ghorbel and Jacot (2011) translate English SentiWordNet entries into French, finding that even if the translation is correct, in some cases two parallel words do not always share the same semantic orientation across both languages due to a difference in common usage. To deal with this issue, Volkova et al. (2013) propose to use crowdsourcing and bootstrapping for learning sentiment lexicons for English, Spanish and Russian from Twitter streams. Gao et al. (2013) found that the use of synonyms and word definitions does not improve the performance of their cotraining approach to learn a Chinese sentiment lexicon from existing sentiment lexicons for English and a corpus of parallel English-Chinese sentences. Chen and Skiena (2014) propose a method for building sentiment lexicons for 136 languages by integrating a variety of linguistic resources to produce a knowledge graph.

2.2. Monolingual sentiment analysis in a multilingual setting

Boiy and Moens (2009) test monolingual classification models for three languages (English, Dutch and French) finding that the French language has the richest vocabulary, while the English language is simpler in terms of vocabulary and syntactic constructions. Cheng and Zhulyn (2012) test two Bayesian classification algorithms on nine languages (English, Dutch, French, Spanish, Italian, Portuguese, German, Chinese and Japanese) concluding that the differences in performance among languages are mainly due to the size of the training set and the length of the test documents. Klinger and Cimiano (2014) perform experiments on English and German, finding the performance values for German to be generally much lower than for English. Severyn et al. (2016) predict the sentiment of YouTube comments written in English and Italian, finding that the performance for Italian was significantly lower than that for English.

Some evaluation campaigns on sentiment analysis dealing with collections in

several languages have been held in recent years. Multilingual Opinion Analysis Task (MOAT) was one of the tasks organised from 2007 to 2010 in the framework of NTCIR-7 and NTCIR-8[1]. Despite its name, the task was not truly multilingual but a combination of five monolingual subtasks for three languages (English, Japanese and Chinese, the latter in both Traditional and Simplified written forms) with an additional Cross-lingual Opinion Question and Answering subtask in NTCIR-8. One of the monolingual subtasks was *Opinion Polarities*, aimed at determining whether the opinion expressed in a sentence was positive, negative or neutral. Participants submitted monolingual results for the languages they chose.

RepLab 2013[2] was one of the labs organized in the framework of CLEF 2013[3]. The goal of the subtask *polarity for reputation classification* was to decide whether the content of tweets written in Spanish or English had positive/negative/neutral implications for a company's reputation (Amigó et al., 2013). Participant systems were not truly multilingual as they considered English and Spanish tweets as separate entities, although in general they extracted the same type of classification features for both languages.

Twitter messages were also considered in (Argueta and Chen, 2014), where polarity classification in English, Spanish and French is performed based on character n-grams and emotion-bearing words and patterns.

2.3. Cross-lingual sentiment analysis

Cross-lingual sentiment analysis consists in using annotated data in a source language (almost always English) to compensate for the lack of labelled data in a target language. One approach consists in training a polarity classifier in English to then apply it to texts written in another language via machine trans-

[1]NII Test Collection for Information Retrieval, `http://research.nii.ac.jp/ntcir/index-en.html`

[2]`http://www.limosine-project.eu/events/replab2013`

[3]Conference and Labs of the Evaluation Forum, formerly known as Cross-Language Evaluation Forum, `http://clef2013.clef-initiative.eu/`

lation (MT). According to Chen and Zhu (2014), text with more sentiment is harder to translate than text with less sentiment. Hiroshi et al. (2004) propose to replace the translation patterns and the bilingual lexicon of classic MT systems with sentiment patterns and a sentiment polarity lexicon. Hajmohammadi et al. (2014) propose to employ both directions of MT simultaneously in order to reduce the effect of MT errors in the classification process. Their experimental results show that classification accuracies vary for different languages, partly due to the fact that MT systems produce translations of varying quality in different languages, and partly due to the disparity in the structure of languages when expressing sentiment information, resulting in sentiment classification showing diverse performance in different languages. In this respect, Demirtas and Pechenizkiy (2013) warn that expanding the training set with new instances taken from a machine-translated corpus does not necessarily increase classification performance, and this is mainly due to the inherent differences in corpora written in different languages. In this regard, they consider that biases due to cultural differences have more impact than inaccurate machine translation techniques.

Balahur and Turchi (2012b) train an SVM classifier for German, Spanish and French data by applying three different MT systems from an English training dataset. Their experiments show that incorrect translations imply an increased amount of features, greater sparseness and more difficulties in identifying a hyperplane which separates the positive and negative examples in the training phase. After manually inspecting the data (Balahur and Turchi, 2012a) they find that the quality of the MT process has implications in the set of features to be used. They conclude in (Balahur and Turchi, 2014) that the gap in classification performance between systems trained on English and translated data is 12% in favor of source language data.

Brooke et al. (2009) adapt English resources and techniques to Spanish, focusing on the modification of their English semantic orientation calculator and the building of Spanish dictionaries. They found that translation seems to have a disruptive effect on previously reliable improvements and that the overall accuracy on translated texts suggests that there is a 5% performance cost for

any automated translation.

Perea-Ortega et al. (2013) obtain a slight improvement in polarity classification performance over an Arabic corpus by considering an English version obtained by means of MT. A similar approach was later tested on Spanish (Martínez Cámara et al., 2014). Wan (2009) proposes to leverage an available English corpus for Chinese sentiment classification by using the English corpus as training data by means of MT. Gui et al. (2013) show that cross-language performance improves when the confidence of the monolingual opinion system is estimated by means of training errors through bilingual transfer self-training and co-training. They also propose a method to improve the transfer of samples during the training phase (Gui et al., 2014).

Balamurali et al. (2012) propose an alternative to MT-based cross-lingual sentiment analysis for languages which do not have an MT system between them but do have WordNets with matching synset identifiers. The main drawback of this technique is the need for automatic word-sense disambiguation, an expensive resource that requires extensive manual annotation, as they report that even low quality word-sense disambiguation leads to an improvement in the performance of sentiment classification.

2.4. Multilingual subjectivity detection and sentiment analysis

Banea et al. (2010) show that multilingual information can improve by almost 5% the performance of subjectivity classification in English (i.e., to determine if a text is objective or subjective). In (Banea et al., 2014) they find that a perfect sense-to-sense mapping between languages is impossible, as a particular sense may denote additional meanings and uses in one language compared to another. However, they also provide evidence that a multilingual feature space is able to rely on double co-occurrence metrics learned from equivalent sense definitions, thus allowing for a more robust modeling than when considering each language individually. Xiao and Guo (2012) confirm on the same dataset that boosting on one view per language improves performance for subjectivity classification with respect to monolingual methods.

Yan et al. (2014) propose a bilingual approach for sentiment analysis consisting in training a single classifier from previously tokenised Chinese and English texts, finding that classification accuracy for English is much better than for Chinese, probably due to the poor quality of word segmentation of Chinese texts.

Davies and Ghahramani (2011) propose a language-independent model for sentiment analysis of Twitter messages, relying on emoticons as unique indicators of sentiment. In the same line, Narr et al. (2012) propose to use emoticons as noisy labels to generate training data from a completely raw set of tweets written in English, German, French and Portuguese, although test data is manually labelled by means of crowdsourcing. They find that a multilingual classifier attains a reasonable performance, although it is worse than the combined accuracies of the monolingual classifiers.

Cui et al. (2011) consider that not only emoticons, but also character and punctuation repetitions are cues of the emotion expressed in a given tweet, independently of the language in which it is written. They propose to construct a graph whose vertices are regular words and emotion tokens while the weight of edges gives a measure of co-ocurrence. They find that the propagation process assigns large positive scores for a majority of tokens, and that negative tweets do not contain many emotion tokens, resulting in a low recall rate on negative tweets, especially for English.

Balahur et al. (2014) translate the English SemEval 2013 Twitter dataset (Chowdhury et al., 2013) into Spanish, Italian, French and German by means of MT systems. Contrary to (Balahur and Turchi, 2012b,a) they find that the use of machine translated data yields similar results to the use of native-speaker translations of the same dataset. Moreover, they find that the use of multilingual data, including those obtained through MT, leads to improved results in sentiment classsification due to the fact that, when using multiple languages to build the classifiers, the features that are relevant are automatically selected, as the feature space becomes sparser. However, they also point out that the performance of the monolingual Spanish sentiment analysis system trained on Spanish

machine translated data can be improved by adding original Spanish data for training (obtained from the Spanish TASS 2013 Twitter dataset (Villena-Román and García-Morera, 2013)) and that even a small number of such texts can lead to a significant increase in classification performance. In contrast, performance decreases when machine-translated English data from SemEval 2013 is used to enlarge the TASS 2013 training corpus for Spanish sentiment analysis (Balahur and Perea-Ortega, 2015).

In contrast to previous work, in this article we present a method for multilingual polarity classification that relies on fusing existing monolingual resources without needing to apply MT techniques, taking as basis the approach we outlined in (Vilares et al., 2015b, 2016a).

3. Building a code-switching corpus

To create the corpus, called 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. For building our resource, we have taken the Spanish-English training set (11 400 tweets). We have filtered out those tweets where all the words belonged to the same language. The resulting collection has a final size of 3 062 tweets. A number of different types of tweets can be found in the corpus:

- Tweets that show (even opposite) sentiment in both languages.
- Tweets where the sentiment is just in the English side of the tweet.
- Tweets where the sentiment is just in the Spanish side of the tweet.
- Tweets where the sentiment relies on language-independent symbols, such as emoticons.

The collection was annotated according to a dual-sentiment scheme, by three speakers fluent in both Spanish and English. In particular, the annotators

assigned each text two scores between 1 and 5: one indicating the positive strength (*ps*) of the tweet and the second one indicating its negative strength (*ns*). This dual scale is usually known as the *SentiStrength scale* (Thelwall et al., 2010). They also were instructed on the Wiebe et al. (2005) annotation guidelines to know how to classify the polarity of a sentence.

For example, *'It was pretty, but too expensive'* would have both a strong positive and negative sentiment. It can also happen that sentiment is expressed by means of a code-mixed expression including English and Spanish words. An example of a sentence presenting this phenomenon in the corpus is *'Im glad we have my tio Crispin fot another year and hopefully diosito le de mucho tiempo mas a nuestro lado'* (*'I'm glad we have my uncle Crispin for another year and hopefully our God will give him much more time by our side'*). Such code-switched expressions are annotated in the same way as their equivalent monolingual expressions, i.e., *'cool fiesta'* would be annotated like *'cool party'* or *'gran fiesta'*.

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. Given the scores of the three annotators, we compute the final strengths of the tweets by averaging the individual positive and negative scores, and rounding to the nearest integer.

There was a total of 200 tweets where the overall sentiment of the sentence was marked as positive by at least one of the annotators and as negative by another one. These can be considered as cases of strong disagreement and they tend to include phenomena such as irony, the occurrence of mixed feelings in the same sentence or the overuse of subjective acronyms. We show below some interesting examples:[4]

[4]To protect the users' privacy, nicknames have been removed. The original code-switching texts are shown as footnotes for clarity reasons. Sentiment scores are indicated as pairs (positive score, negative score).

- *'Talking about the devil,.,., my mommy just arrived :)'*.[5] The sentiment scores assigned to the tweet were: (2,1), (1,3) and (1,3).

- *'This movie is badass like damn and makes me cry lol'*.[6] In particular, the tweet was scored with (4,1), (1,5) and (1,4).

- *'lol miss you too!!! :p mmmmm hahahaha'*.[7] The tweet was assigned the following individual scores: (4,1), (1,2) and (4,1).

Positive	**%tweets**	**Negative**	**%tweets**
1	63.3	1	69.4
2	26.6	2	19.6
3	7.5	3	8.4
4	2.4	4	2.2
5	0.3	5	0.1

Table 1: Frequency distribution of the SentiStrength scores on the EN-ES-CS CORPUS

Table 1 shows the frequency distribution of the SentiStrength scores and how annotators often tend to find slight levels of subjectivity, while highly subjective tweets tend to be less frequent.[8]

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

Table 2: 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)

[5] *'Hablando del demonio,.,., ya llego mi mommy :)'*.

[6] *'This movie is badass like damm me ase llorar lol'*.

[7] *'lol miss you too!!! :p mmmmm jajajaja'*

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

The results are coherent with other corpora annotated according to these criteria (Thelwall et al., 2010; Vilares et al., 2015d). 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 by downloading tweets posted by people from Texas and California, where English is the primary language. Table 2 reflects these particularities. [9] In total, our collection contains 24 758 English terms, with 5 565 unique words, of which 3 576 turned out to be out-of-vocabulary (OOV). 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 cues was also sometimes used 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.

3.1. Trinary scale conversion

A second labelling 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; Rosenthal et al., 2014a, 2015). If positive strength is greater than negative strength, the tweet was considered *positive*. If negative strength is greater than the positive one, the tweet was considered *negative*. Otherwise, it was taken

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

as *neutral*.[10] After the conversion, we obtained a collection where the *positive* class represents 31.45% of the corpus and the *negative* one 25.67%, the remaining 42.88% of tweets being neutral. We used this annotation for the experiments reported in the following sections.

4. A multilingual sentiment analysis model

As explained, our goal is to compare the performance of supervised monolingual models based on *bag-of-words*, often used in SA tasks, with respect to their corresponding multilingual version (i.e. a model that is a collection of weights from English and Spanish features). To do this, we rely on standard sets of features. The aim of this article is not to introduce a new sentiment analysis architecture, but to show how current state-of-the-art supervised approaches can successfully address (or not) situations where monolingual, multilingual and code-switching texts appear. We relied on an L2-regularised logistic regression (Fan et al., 2008). In general, linear classifiers have provided state-of-the-art performance since early research on SA (Pang et al., 2002; Paltoglou and Thelwall, 2010; Mohammad et al., 2013) and in particular, logistic regression is a good fit for this task (Jurafsky and Martin, 2016).

4.1. Basic features

Four atomic sets of features are considered:

- *Words (W)*: Simple statistical model that counts the frequencies of words in a text.
- *Lemmas (L)*: Each term is lemmatised to reduce sparsity, using lexicon-based methods that rely on the Ancora corpus (Taulé et al., 2008) for Spanish and Multext (Ide and Véronis, 1994) and a set of rules[11] for

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

[11] `http://sourceforge.net/p/zpar/code/HEAD/tree/src/english/morph/aux_lexicon.cpp`

English.

- *Psychometric properties (P)*: Emotions, psychological concepts (e.g. *anger*) or topics (e.g. *job*) that commonly appear in messages. We rely on the LIWC dictionaries (Pennebaker et al., 2001) to detect these.

	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			

Table 3: Performance of taggers on a code-switching sentence from Twitter: *adverb* (ADV), *adjective* (ADJ), *prepositions* and *postpositions* (ADP), *determiner* (DET), *noun* (NOUN), *pronoun* (PRON), *verb* (VERB) and *other category* (X). The corresponding English sentence is: *'We are working hard on putting available the best products of Spain, thank you'*

- *Part-of-speech tags (T)*: The grammatical categories were obtained using the Stanford Maximum Entropy model (Toutanova and Manning, 2000). We trained an *en* and an *es* tagger using the Google universal PoS tagset (Petrov et al., 2011) and joined the Spanish and English corpora to train a combined *en-es* tagger. The aim was to build a model that does not need any language detection to tag samples written in different languages, or even code-switching sentences. Table 3 shows how the three taggers work on a real code-switching sentence from Twitter, illustrating how the *en-es* tagger effectively tackles them. The accuracy of the *en* and *es* taggers was 98.12%[12] and 96.03% respectively. The multilingual tagger obtained 98.00% and 95.88% over the monolingual test sets.

[12]Note that Toutanova and Manning reported 97.97% on the Penn Treebank tagset, which is bigger than the Google Universal tagset (48 vs 12 tags).

These atomic sets of features can be combined to obtain a rich linguistic model that improves performance (Section 5).

4.2. Syntactic features

We also consider syntactic dependencies between words as features. Dependency parsing is defined as the process of obtaining a dependency tree for a given sentence. Let $S = [s_1 s_2 ... s_{n-1} s_n]$ be a sentence[13] of length n, where s_i indicates the token at the i^{th} position; a *dependency tree* is a labelled directed graph with edges of the form (s_j, m_{jk}, s_k). Each such edge represents a binary relation (*dependency*) between two words, called the *head* (s_j) and *dependent* (s_k) tokens, and the kind of syntactic relation (such as subject, object, etc.) is described by the label m_{jk}.

To obtain such trees, we trained an *en*, *es* and an *en-es* parser (Vilares et al., 2016b) using MaltParser (Nivre et al., 2007). In order to obtain competitive results for each specific language, we relied on MaltOptimizer (Ballesteros and Nivre, 2012). The parsers were trained on the Universal Dependency Treebanks v2.0 (McDonald et al., 2013) and evaluated against the monolingual test sets. The Labelled Attachment Score (LAS)[14] of the Spanish and English monolingual parsers was 80.54% and 88.35%, respectively. The multilingual model achieved an LAS of 78.78% and 88.65% (the latter implies a significant improvement with respect to the monolingual model, using Bikel's randomised parsing evaluation comparator and $p < 0.05$). Figure 1 shows an example of how the *en*, *es* and *en-es* parsers work on a code-switching sentence.

In the next step, words, lemmas, psychometric properties and PoS tags are used to extract *enriched generalised triplet* features (Vilares et al., 2015a). Let (s_j, m_{ij}, s_k) be a triplet with $s_j, s_k \in W$ and generalisation functions, $g_1, g_2 : W \rightarrow W \cup L \cup P \cup T$, a *generalised triplet* is defined as $(g_1(s_j), m_{ij}, g_2(s_k))$.

[13] An artificial token s_0, named ROOT, is usually added for technical reasons.

[14] The LAS metric measures the proportion of words that are assigned both the correct head and the correct dependency label by the parser.

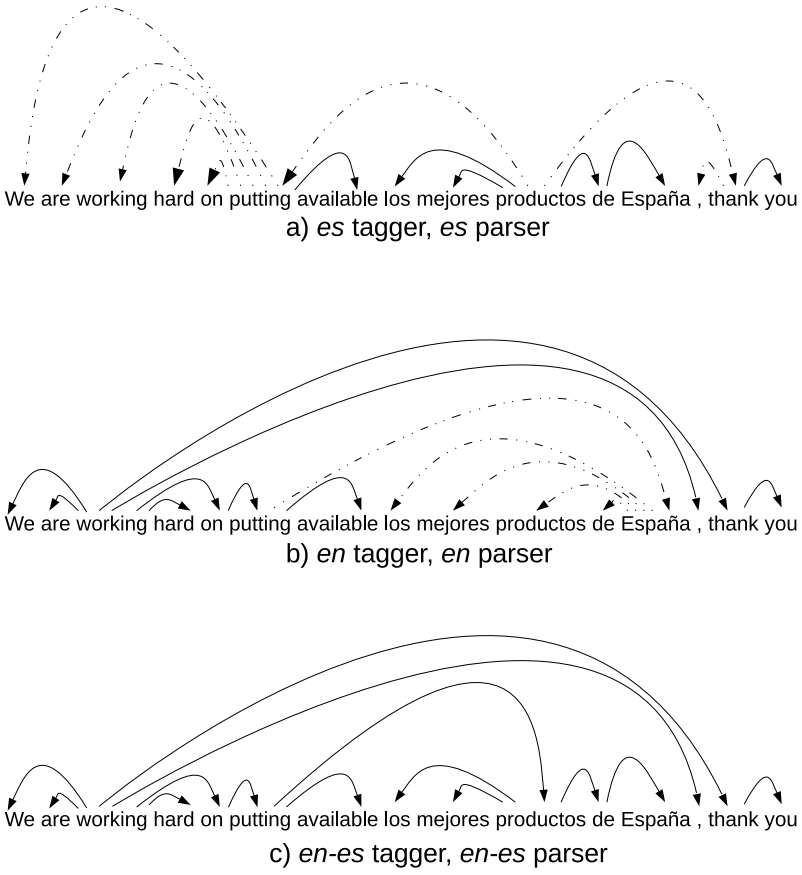

Figure 1: Example of a tweet parsed with the *en*, *es* and *en-es* dependency parsers. Dotted/dashed lines represent incorrectly-parsed dependencies.

4.3. N-gram features

N-gram features capture shallow structure of sentences, identifying local relations between words (e.g. *'not good'* becomes *'not_good'*). In particular, we are considering bi-grams of words, lemmas and psychometric properties (e.g. *'not good'* would become *'negation_positive-emotion'*).

5. Experimental framework

We test three different approaches in our experiments:

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 message 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.

These approaches are evaluated on standard monolingual corpora, taking accuracy and F1-measure as the reference metrics. The monolingual collections

are then joined to create a multilingual corpus, which helps us compare the performance of the approaches when tweets come from two different languages. An evaluation over a code-switching test set is also carried out. Concretely, the following corpora have been used:

1. *SemEval 2014 task B* corpus (Rosenthal et al., 2014a): A set of English tweets[15] split into training (8 200 tweets), development (1 416) and test sets[16] (5 752). Each tweet was manually classified as *positive*, *none* or *negative*.
2. *TASS 2014* 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 labelled tweets, named *1K test set*. The tweets are labelled with *positive*, *none*, *negative* and *mixed*, but in this study the *mixed* class was treated as *none*, following the same criteria as in SemEval 2014.
3. Multilingual corpora resulting from merging SemEval 2014 and TASS 2014 corpora. These two test sets were merged to create two synthetic multilingual corpora: (1) SemEval 2014 + TASS 2014 1K (English is the majority language) and (2) SemEval 2014 + TASS 2014 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 English-Spanish code-switching corpus described in Sect. 3.

[15]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.

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

6. Experimental results

We show below the performance of each model in each of the four proposed configurations: (1) an English monolingual corpus, (2) a Spanish monolingual corpus, (3) a multilingual corpus which combines the two monolingual collections and (4) the code-switching (Spanish-English) corpus presented in Sect. 3.

Features	**F1-measure**			**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	59.1	59.1	**60.2**	61.0	61.00	**61.5**
Bigrams of L	**59.9**	**59.9**	**59.9**	**61.8**	**61.8**	61.3
Bigrams of P	**60.6**	**60.6**	59.8	**61.3**	**61.3**	60.4
Triplets of W	53.1	53.1	**55.8**	56.4	56.4	**57.8**
Triplets of L	56.0	56.0	**57.2**	58.7	58.7	**59.2**
Triplets of P	**57.4**	**57.4**	56.9	**58.3**	58.2	57.6
Combined (W,P,T)	68.0	**69.0**	68.2	68.5	**68.6**	**68.6**
Combined (L,P,T)	**68.0**	67.8	67.9	**68.4**	**68.4**	68.3
Combined (W,P)	68.2	**68.3**	68.1	**68.7**	**68.7**	68.5
Combined (L,P)	**68.0**	**68.0**	67.8	**68.6**	68.5	68.3

Table 4: Performance (%) on the SemEval 2014 test set. 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.

Table 4 shows the performance of the three models on the SemEval English monolingual test set. With respect to the evaluation on the Spanish monolingual corpora, results on the TASS 2014 corpora are shown in Table 5, including results on both the general and the TASS 2014-1K test sets. Table 6 shows the performance both of the multilingual approach and the monolingual pipeline with language detection when analysing texts in different languages. Finally,

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
Words (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**
Lemmas (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**
Psychometric (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**
PoS-tags (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
Bigrams of W	**54.4**	54.2	53.9	**52.9**	52.7	52.1	58.2	58.3	**58.9**	58.4	58.4	**58.7**
Bigrams of L	**55.5**	**55.4**	54.3	**54.0**	53.9	52.2	58.6	58.6	**59.3**	58.7	58.7	**59.3**
Bigrams of P	47.6	47.6	**48.7**	46.0	46.0	**47.0**	51.3	51.2	**53.2**	51.3	51.3	**53.2**
Triplets of 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**
Triplets of 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**
Triplets of 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**
Combined (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
Combined (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
Combined (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**
Combined (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 5: Performance (%) on the TASS test sets. 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 7 shows the performance of the three proposed approaches on the code-switching test set.

7. Discussion

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 appear. 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. Vilares et al. (2015a) explain how features of this kind become useful when the training data becomes larger.

Features	SemEval+TASS-1K				SemEval+TASS-general			
	F1		Accuracy		F1		Accuracy	
	pipe	en-es	pipe	en-es	pipe	en-es	pipe	en-es
Words (W)	**64.5**	63.7	**64.9**	64.2	64.3	**64.5**	64.6	**64.7**
Lemmas (L)	**64.5**	**64.5**	**65.0**	64.8	64.3	**64.4**	**64.7**	**64.7**
Psychometric (P)	**60.5**	59.1	**61.2**	60.0	58.7	**59.4**	59.0	**59.7**
PoS-tags (T)	48.1	**49.2**	**51.3**	50.2	**49.2**	46.2	**49.7**	48.1
Bigrams of W	58.3	**59.2**	59.6	**59.8**	58.3	**59.0**	58.6	**58.9**
Bigrams of L	**59.2**	59.0	**60.4**	59.7	58.7	**59.4**	59.0	**59.5**
Bigrams of P	58.6	**58.8**	**58.7**	58.1	52.0	**53.8**	52.2	**53.9**
Triplets of W	53.1	**54.4**	**55.7**	55.5	54.1	**54.9**	54.6	**55.2**
Triplets of L	**55.9**	55.8	**57.9**	56.9	55.9	**56.5**	56.3	**56.6**
Triplets of P	**55.8**	55.5	**56.1**	55.8	50.6	**52.7**	50.3	**52.8**
Combined (W,P,T)	**67.8**	67.0	**67.1**	66.9	66.2	**66.3**	**66.5**	**66.5**
Combined (L,P,T)	**67.0**	66.8	**67.2**	66.8	65.9	**66.1**	**66.3**	**66.3**
Combined (W,P)	**67.1**	67.0	**67.1**	67.0	66.1	**66.2**	66.4	**66.5**
Combined (L,P)	**66.9**	66.7	**67.0**	66.8	65.8	65.9	**66.1**	**66.1**

Table 6: Performance (%) on the multilingual test set. 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.

7.1. English corpus

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.

	F1-measure				Accuracy			
Features	**en**	**es**	**pipe**	**en-es**	**en**	**es**	**pipe**	**en-es**
Words (W)	**54.2**	45.2	51.6	54.1	**55.7**	47.7	52.7	54.89
Lemmas (L)	54.3	46.2	51.9	**55.7**	55.9	48.9	53.0	**56.4**
Psychometric (P)	52.2	40.8	50.0	**53.3**	53.0	43.6	50.7	**53.7**
PoS-tags (T)	38.5	34.4	40.2	**39.6**	**45.1**	39.3	44.7	43.2
Bigrams of W	49.3	45.1	48.5	**51.9**	54.3	47.5	51.7	**54.3**
Bigrams of L	50.1	46.4	49.1	**51.4**	**55.0**	48.9	52.2	53.6
Bigrams of P	**47.7**	37.3	45.2	46.8	**49.5**	40.5	46.1	46.9
Triplets of W	46.6	30.2	43.1	**47.1**	**52.6**	36.5	46.0	50.7
Triplets of L	47.4	42.4	45.6	**47.8**	**53.0**	44.7	49.0	50.4
Triplets of P	**46.2**	36.2	44.5	45.6	**48.1**	40.6	45.7	46.0
Combined (W,P,T)	58.3	47.1	56.1	**58.5**	**59.2**	48.3	56.5	58.5
Combined (L,P,T)	57.7	48.9	55.6	**58.6**	58.6	49.7	56.1	**59.1**
Combined (W,P)	58.0	48.4	55.9	**58.8**	58.7	49.9	56.4	**58.8**
Combined (L,P)	58.2	49.3	55.6	**58.9**	58.9	50.8	56.1	**59.3**

Table 7: Performance (%) on the code-switching set. For each row, the best values of F1 and accuracy are shown in boldface.

7.2. Spanish corpora

With respect to the evaluation on the TASS 2014 and TASS 2014-1k corpora the *es* model obtains the best results, followed by the *pipe* and the *en-es* models. In the TASS 2014-1k test set, the language detection system misclassified 17 of the manually labelled tweets, and the impact of the monolingual model with language detection is also small. Results obtained on the TASS 2014 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 sug-

gests that multilingual models present advantages when messages in different languages need to be analysed.

7.3. Synthetic multilingual corpus

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.

7.4. Code-switching corpus

The accuracy obtained by the proposed models on this corpus is lower than on the monolingual corpora. This suggests that analysing 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 lemmatisation, 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 article is more robust on environments containing code-switching tweets and tweets in different languages. The *es* model performs poorly, 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 hypothesise this is due to the

higher presence of English phrases, which made it possible to extract the sentiment of the texts in many cases.

8. Conclusion

In this article, 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 analysed and (3) where code-switching texts appear. To evaluate scenario (3), we have presented together with this article 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.

As future work, we would like to evaluate deep learning architectures such as convolutional neural networks on code-switching texts, since we think their ability to exploit spatially-local correlations can be helpful for the purpose at hand: we observed that code-switching sentences tend to contain continuous slices of text written in the same language, which can be relevant to determine

the global sentiment of the sentence. Additionally, we plan to explore the performance of multilingual unsupervised approaches on this kind of environments. Current multilingual supervised approaches require labeled data to be trained, which is not always available especially when the target languages are scarce in resources.

9. Acknowledgments

This research was supported by the Ministerio de Economía y Competitividad (FFI2014-51978-C2) and Xunta de Galicia (R2014/034). 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).

References

Aldayel, H.K., Azmi, A.M., 2015. Arabic tweets sentiment analysis — a hybrid scheme. Journal of Information Science

Amigó, E., de Albornoz, J.C., Chugur, I., Corujo, A., Gonzalo, J., Martín-Wanton, T., Meij, E., de Rijke, M., Spina, D., 2013. Overview of RepLab 2013: Evaluating online reputation monitoring systems, in: Forner, P., Müller, H., Paredes, R., Rosso, P., Stein, B. (Eds.), Information Access Evaluation. Multilinguality, Multimodality, and Visualization — 4th International Conference of the CLEF Initiative, CLEF 2013, Valencia, Spain, September 23-26, 2013. Proceedings. Springer, Berlin and Heidelberg. volume 8138 of *Lecture Notes in Computer Science*, pp. 333–352.

Arakawa, Y., Kameda, A., Aizawa, A., Suzuki, T., 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, 1416–1423.

Argueta, C., Chen, Y., 2014. Multi-Lingual Sentiment Analysis of Social Data Based on Emotion-Bearing Patterns, in: Lin, S.d., Ku, L.W., Cambria, E., Kuo, T.T. (Eds.), SocialNLP 2014. The Second Workshop on Natural Language Processing for Social Media in conjunction with COLING-2014. Proceedings of the Workshop, Dublin, Ireland. pp. 38–43.

Baccianella, S., Esuli, A., Sebastiani, F., 2010. SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining, in: Chair, N.C.C., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Rosner, M., Tapias, D. (Eds.), Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10), European Language Resources Association (ELRA), Valletta, Malta.

Balahur, A., Perea-Ortega, J.M., 2015. Sentiment analysis system adaptation for multilingual processing: The case of tweets. Information Processing and Management 51, 547–556.

Balahur, A., Turchi, M., 2012a. Comparative Experiments for Multilingual Sentiment Analysis Using Machine Translation, in: Gaber, M.M., Cocea, M., Weibelzahl, S., Menasalvas, E., Labbe, C. (Eds.), SDAD 2012, The 1st International Workshop on Sentiment Discovery from Affective Data, Bristol, UK. pp. 75–86.

Balahur, A., Turchi, M., 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.

Balahur, A., Turchi, M., 2014. Comparative experiments using supervised learning and machine translation for multilingual sentiment analysis. Computer Speech and Language 28, 56–75.

Balahur, A., Turchi, M., Steinberger, R., Perea-Ortega, J.M., Jacquet, G., Kucuk, D., Zavarella, V., Ghali, A.E., 2014. Resource Creation and Evaluation

for Multilingual Sentiment Analysis in Social Media Texts, in: Chair, N.C.C., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., Piperidis, S. (Eds.), Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), European Language Resources Association (ELRA), Reykjavik, Iceland.

Balamurali, A.R., Joshi, A., Bhattacharyya, P., 2012. Cross-lingual sentiment analysis for Indian languages using linked wordnets, in: Kay, M., Boitet, C. (Eds.), COLING 2012. 24th International Conference on Computational Linguistics. Proceedings of COLING 2012: Posters, Mumbai, India. pp. 73–81.

Balazs, J.A., Velásquez, J.D., 2016. Opinion Mining and Information Fusion: A Survey. Information Fusion 27, 95–110. .

Ballesteros, M., Nivre, J., 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, C., Mihalcea, R., Wiebe, J., 2010. Multilingual Subjectivity: Are More Languages Better?, in: Huang, C.R., Jurafsky, D. (Eds.), COLING 2010. 23rd International Conference on Computational Linguistics. Proceedings of the Conference, Tsinghua University Press, Beijing. pp. 28–36.

Banea, C., Mihalceaa, R., Wiebe, J., 2014. Sense-level subjectivity in a multilingual setting. Computer Speech & Language 28, 7–19.

Boiy, E., Moens, M., 2009. A machine learning approach to sentiment analysis in multilingual Web texts. Information Retrieval 12, 526–558.

Brooke, J., Tofiloski, M., Taboada, M., 2009. Cross-linguistic sentiment analysis: From English to Spanish, in: Proceedings of RANLP 2009, Recent Advances in Natural Language Processing, Bovorets, Bulgaria. pp. 50–54.

Cambria, E., Rajagopal, D., Olsher, D., Das, D., 2013a. Big social data analysis. Big data computing , 401–414, 2013.

Cambria, E., Schuller, B., Liu, B., Wang, H., Havasi, C., 2013b. Knowledge-based approaches to concept-level sentiment analysis. IEEE Intelligent Systems , 12–14.

Chen, B., Zhu, X., 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, Y., Skiena, S., 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, A., Zhulyn, O., 2012. A system for multilingual sentiment learning on large data sets, in: Kay, M., Boitet, C. (Eds.), COLING 2012. 24th International Conference on Computational Linguistics. Proceedings of COLING 2012: Technical Papers, Mumbai, India. pp. 577–592.

Chowdhury, M.F.M., Guerini, M., Tonelli, S., Lavelli, A., 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.

Cotelo, J., Cruz, F., Enríquez, F., Troyano, J., 2016. Tweet categorization by combining content and structural knowledge. Information Fusion 31, 54–64. URL: http://linkinghub.elsevier.com/retrieve/pii/S1566253516000099, doi:10.1016/j.inffus.2016.01.002.

Cruz, F., Troyano, J.A., Pontes, B., Ortega, F.J., 2014. Building layered, mul-

tilingual sentiment lexicons at synset and lemma levels. Expert Systems with Applications 41, 5984–5994.

Cui, A., Zhang, M., Liu, Y., Ma, S., 2011. Emotion Tokens: Bridging the Gap Among Multilingual Twitter Sentiment Analysis, in: Salem, M.V.M., Shaalan, K., Oroumchian, F., Shakery, A., Khelalfa, H. (Eds.), Information Retrieval Technology. 7th Asia Information Retrieval Societies Conference, AIRS 2011, Dubai, United Arab Emirates, December 18-20, 2011. Proceedings. Springer, Berlin and Heidelberg. volume 7097 of *Lecture Notes in Computer Science*, pp. 238–249.

Dang, Y., Zhang, Y., Hu, P.J., Brown, S.A., Ku, Y., Wang, J., Chen, H., 2014. An integrated framework for analyzing multilingual content in Web 2.0 social media. Decision Support Systems 61, 126–135.

Davies, A., Ghahramani, Z., 2011. Language-independent Bayesian sentiment mining of Twitter, in: The 5th SNA-KDD Workshop'11 (SNA-KDD'11), ACM, San Diego, CA.

Demirtas, E., Pechenizkiy, M., 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), ACM, Chicago, IL, USA. p. Article No. 9.

Fan, R.E., Chang, K.W., Hsieh, C.J., Wang, X.R., Lin, C.J., 2008. LIBLINEAR: A library for large linear classification. The Journal of Machine Learning Research 9, 1871–1874.

Gao, D., Wei, F., Li, W., Liu, X., Zhou, M., 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.

Ghorbel, H., Jacot, D., 2011. Sentiment analysis of French movie reviews, in: Pallotta, V., Soro, A., Vargiu, E. (Eds.), Advances in Distributed Agent-

Based Retrieval Tools. Springer, Berlin and Heidelberg. volume 361 of *Studies in Computational Intelligence*, pp. 97–108.

Gonzalez-Agirre, A., Laparra, E., Rigau, G., 2012. Multilingual Central Repository version 3.0, in: Chair, N.C.C., Choukri, K., Declerck, T., Doğan, M.U., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., Piperidis, S. (Eds.), Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12), European Language Resources Association (ELRA), Istanbul, Turkey.

Gui, L., Xu, R., Lu, Q., Xu, J., Xu, J., Liu, B., Wang, X., 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.

Gui, L., Xu, R., Xu, J., Yuan, L., Yao, Y., Zhou, J., Qiu, Q., Wang, S., Wong, K.F., Cheung, R., 2013. A mixed model for cross lingual opinion analysis, in: Zhou, G., Li, J., Zhao, D., Feng, Y. (Eds.), Natural Language Processing and Chinese Computing, Springer, Heidelberg, NewYork, Dordrecht and London. pp. 93–104.

Habernal, I., Ptáček, T., Steinberg, J., 2014. Supervised sentiment analysis in Czech social media. Information Processing and Management 50, 693–707.

Hajmohammadi, M.S., Ibrahim, R., Selamat, A., 2014. Bi-view semi-supervised active learning for cross-lingual sentiment classification. Information Processing and Management 50, 718–732.

Hayes, A., Krippendorff, K., 2007. Answering the call for a standard reliability measure for coding data. Communication Methods and Measures 1, 77–89.

Hiroshi, K., Tetsuya, N., Hideo, W., 2004. Deeper sentiment analysis using machine translation technology, in: Proceedings of the 20th international

conference on Computational Linguistics (COLING 2004), Geneva, Switzerland.

Hogenboom, A., Heerschop, B., Frasincar, F., Kaymak, U., de Jong, F., 2014. Multi-lingual support for lexicon-based sentiment analysis guided by semantics. Decision Support Systems 62, 43–53.

Ide, N., Véronis, J., 1994. Multext: Multilingual text tools and corpora, in: Proceedings of the 15th conference on Computational linguistics-Volume 1, Association for Computational Linguistics. pp. 588–592.

Inrak, P., Sinthupinyo, S., 2010. Applying latent semantic analysis to classify emotions in Thai text, in: 2nd International Conference on Computer Engineering and Technology (ICCET), IEEE, Chengdu. pp. 450–454.

Jurafsky, D., Martin, J.H., 2016. Classification: Naive Bayes, Logistic Regression, Sentiment. Chapter 7 of Speech and Language Processing (3rd ed. draft).

Khaleghi, B., Khamis, A., Karray, F.O., Razavi, S.N., 2013. Multisensor data fusion: A review of the state-of-the-art. Information Fusion 14, 28–44.

Kim, J., Jung, H.Y., Nam, S.H., Lee, Y., Lee, J.H., 2009. Found in translation: Conveying subjectivity of a lexicon of one language into another using a bilingual dictionary and a link analysis algorithm, in: Li, W., Mollá-Aliod, D. (Eds.), Computer Processing of Oriental Languages. Language Technology for the Knowledge-based Economy. Springer, Berlin and Heidelberg. volume 5459 of *Lecture Notes in Computer Science*, pp. 112–121.

Klinger, R., Cimiano, P., 2014. The USAGE review corpus for fine-grained, multi-lingual opinion analysis, in: Chair, N.C.C., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., Piperidis, S. (Eds.), Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), European Language Resources Association (ELRA), Reykjavik, Iceland.

Lui, M., Baldwin, T., 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.

Martínez Cámara, E., Martín Valdivia, M.T., Molina-González, M.D., Perea-Ortega, J.M., 2014. Integrating Spanish lexical resources by meta-classifiers for polarity classification. Journal of Information Science 40, 538–554.

McDonald, R., Nivre, J., Quirmbach-Brundage, Y., Goldberg, Y., Das, D., Ganchev, K., Hall, K., Petrov, S., Zhang, H., Täckström, O., Bedini, C., Castelló, N., Lee, J., 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.

Medagoda, N., Shanmuganathan, S., Whalley, J., 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, IEEE, Colombo, Sri Lanka.

Mohammad, S.M., Kiritchenko, S., Zhu, X., 2013. NRC-Canada: Building the state-of-the-art in sentiment analysis of tweets. arXiv preprint arXiv:1308.6242 .

Nakov, P., Rosenthal, S., Kozareva, Z., Stoyanov, V., Ritter, A., Wilson, T., 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.

Narr, S., Hülfenhaus, M., Alnayrak, S., 2012. Language-Independent Twitter Sentiment Analysis, in: Workshop on Knowledge Discovery, Data Mining and Machine Learning (KDML-2012), Dortmund, Germany.

Neri, F., Aliprandi, C., Capeci, F., Cuadros, M., By, T., 2012. Sentiment analysis on social media, in: Proceedings of the 2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, IEEE. pp. 951–958.

Nivre, J., Hall, J., Nilsson, J., Chanev, A., Eryigit, G., Kübler, S., Marinov, S., Marsi, E., 2007. MaltParser: A language-independent system for data-driven dependency parsing. Natural Language Engineering 13, 95–135. URL: `http://dblp.uni-trier.de/db/journals/nle/nle13.html#NivreHNCEKMM07`.

Paltoglou, G., Thelwall, M., 2010. A study of information retrieval weighting schemes for sentiment analysis, in: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics. pp. 1386–1395.

Pang, B., Lee, L., Vaithyanathan, S., 2002. Thumbs up? Sentiment classification using machine learning techniques, in: Proceedings of EMNLP, pp. 79–86.

Pennebaker, J.W., Francis, M.E., Booth, R.J., 2001. Linguistic inquiry and word count: LIWC 2001. Mahway: Lawrence Erlbaum Associates , 71.

Perea-Ortega, J.M., Martín-Valdivia, M.T., Ureña López, L.A., Martínez-Cámara, E., 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, 1864–1877.

Petrov, S., Das, D., McDonald, R., 2011. A universal part-of-speech tagset. arXiv preprint arXiv:1104.2086 .

Román, J., Martínez-Cámara, E., García-Morera, J., Jiménez-Zafra, S.M., 2015. TASS 2014-The Challenge of Aspect-based Sentiment Analysis. Procesamiento del Lenguaje Natural 54, 61–68.

Rosenthal, S., Nakov, P., Kiritchenko, S., Mohammad, S.M., Ritter, A., Stoyanov, V., 2015. Semeval-2015 task 10: Sentiment analysis in Twitter, in:

Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015).

Rosenthal, S., Nakov, P., Ritter, A., Stoyanov, V., 2014a. Semeval-2014 task 9: Sentiment analysis in Twitter, in: Proceedings of The 8th International-Workshop on Semantic Evaluation (SemEval 2014), pp. 411–415.

Scholz, T., Conrad, S., 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. Springer, Berlin and Heidelberg. volume 7934 of *Lecture Notes in Computer Science*, pp. 272–277.

Severyn, A., Moschitti, A., Uryupina, O., Plank, B., Filippova, K., 2016. Multilingual opinion mining on YouTube. Information Processing and Management 52, 46–60.

Solorio, T., Blair, E., Maharjan, S., Bethard, S., Diab, M., Gohneim, M., Hawwari, A., AlGhamdi, F., Hirschberg, J., Chang, A., Fung, P., 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, Association for Computational Linguistics, Doha, Qatar. pp. 62–72.

Steinberger, R., 2012. A survey of methods to easy the development of highly multilingual text mining applications. Language Resources and Evaluation 46, 155–176.

Taulé, M., Martí, M.A., Recasens, M., 2008. AnCora: Multilevel Annotated Corpora for Catalan and Spanish, in: Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odjik, J., Piperidis, S., Tapias, D. (Eds.), Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08), Marrakech, Morocco. pp. 96–101.

Thelwall, M., Buckley, K., Paltoglou, G., 2011. Sentiment in Twitter events. J. Am. Soc. Inf. Sci. Technol. 62, 406–418. URL: http://dx.doi.org/10.1002/asi.21462, doi:10.1002/asi.21462.

Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., Kappas, A., 2010. Sentiment Strength Detection in Short Informal Text. Journal of the American Society for Information Science and Technology 61, 2544–2558.

Toutanova, K., Manning, C.D., 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.

Vilares, D., Alonso, M.A., Gómez-Rodríguez, C., 2015a. On the usefulness of lexical and syntactic processing in polarity classification of Twitter messages. Journal of the Association for Information Science Science and Technology 66, 1799–1816. doi:10.1002/asi.23284.

Vilares, D., Alonso, M.A., Gómez-Rodríguez, C., 2015b. 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, Association for Computational Linguistics, Lisboa, Portugal. pp. 2–8. URL: http://aclweb.org/anthology/W15-2902.

Vilares, D., Alonso, M.A., Gómez-Rodríguez, C., 2015c. A syntactic approach for opinion mining on Spanish reviews. Natural Language Engineering 21, 139–163.

Vilares, D., Alonso, M.A., Gómez-Rodríguez, C., 2016a. EN-ES-CS: An English-Spanish Code-Switching Twitter Corpus for Multilingual Sentiment Analysis, in: Proceedings of the 10th edition of the Language Resources and Evaluation Conference (LREC 2016), ELRA, Portoroz, Eslovenia. pp. 4149–4153.

Vilares, D., Gómez-Rodríguez, C., Alonso, M.A., 2016b. One model, two languages: training bilingual parsers with harmonized treebanks, in: Proc. of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016), p. 425–431.

Vilares, D., Thelwall, M., Alonso, M.A., 2015d. The megaphone of the people? Spanish SentiStrength for real-time analysis of political tweets. Journal of Information Science to appear, 799–813.

Villena-Román, J., García-Morera, J., 2013. TASS 2013 — workshop on sentiment analysis at SEPLN 2013: An overview, in: Díaz Esteban, A., Alegría Loinaz, I.n., Villena Román, J. (Eds.), XXIX Congreso de la Sociedad Española de Procesamiento de Lenguaje Natural (SEPLN 2013). TASS 2013 - Workshop on Sentiment Analysis at SEPLN 2013, Madrid, Spain. pp. 112–125.

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

Volkova, S., Wilson, T., Yarowsky, D., 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, ACL, Sofia, Bulgaria. pp. 505–510.

Vossen, P. (Ed.), 1998. EuroWordNet. A Multilingual Database with Lexical Semantic Networks. Kluwer Academic Publishers, Dordrecht, The Netherlands. Reprinted from Computers and the Humanities, Volume 32, Nos. 2–3, 1998.

Wan, X., 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, ACL. World Scientific Publishing Co Pte Ltd, Suntec, Singapore. pp. 235–243.

Wiebe, J., Wilson, T., Cardie, C., 2005. Annotating expressions of opinions and emotions in language. Language resources and evaluation 39, 165–210.

Xiao, M., Guo, Y., 2012. Multi-View AdaBoost for Multilingual Subjectivity Analysis, in: Kay, M., Boitet, C. (Eds.), COLING 2012. 24th International Conference on Computational Linguistics. Proceedings of COLING 2012: Technical Papers, Mumbai, India. pp. 2851–2866.

Yan, G., He, W., Shen, J., Tang, C., 2014. A bilingual approach for conducting Chinese and English social media sentiment analysis. Computer Networks 75, 491–503.

Zhang, C., Zeng, D., Li, J., Wang, F.Y., Zuo, W., 2009. Sentiment analysis of Chinese documents: From sentence to document level. Journal of the American Society for Information Science and Technology 60, 2474–2487.