Sentiment Analysis on Monolingual, Multilingual and Code-Switching Twitter Corpora

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
{david.vilares, miguel.alonso, carlos.gomez}@udc.es

Abstract

We address the problem of performing polarity classification on Twitter over different languages, focusing on English and Spanish, comparing three techniques: (1) a monolingual model which knows the language in which the opinion is written, (2) a monolingual model that acts based on the decision provided by a language identification tool and (3) a multilingual model trained on a multilingual dataset that does not need any language recognition step. Results show that multilingual models are even able to outperform the monolingual models on some monolingual sets. We introduce the first code-switching corpus with sentiment labels, showing the robustness of a multilingual approach.

1 Introduction

Noisy social media, such as Twitter, are especially interesting for *sentiment analysis* (SA) and *polarity classification* tasks, given the amount of data and their popularity in different countries, where users simultaneously publish opinions about the same topic in different languages (Cambria et al., 2013a; Cambria et al., 2013b). Some expressions are written in different languages, making the polarity classification harder. In this context, handling texts in different languages becomes a real need. We evaluate three machine learning models, considering Spanish (*es*), English (*en*) and its multilingual version, English-Spanish (*en-es*):

1. *Multilingual approach (en-es model)*: A model does not need to recognise the language of the text. The *en* and *es* training and development corpora are merged to train an unique *en-es* sentiment classifier.
2. *Monolingual approach (en and es models)*: The ideal case where the language of the text is known and the right model is executed. Each language model is trained and tuned on a monolingual corpus.
3. *Monolingual pipeline with language detection (pipe model)*: Given an unknown text, we first identify the language of the message through `lang.py` (Lui and Baldwin, 2012). The output language set was constrained to Spanish and English to make sure every tweet is classified and guarantee a fair comparison with the rest of the approaches. The training was done in the same way as in the monolingual approach, as we know the language of the texts. `Lang.py` is just needed for evaluation. The language is predicted, the corresponding monolingual classifier is called and the outputs are joined to compare them to the gold standard.

The approaches are evaluated on: (1) an English monolingual corpus, (2) a Spanish monolingual corpus (3) a multilingual corpus which combines the two monolingual collections and (4) a code-switching (Spanish-English) corpus, that is introduced together with this paper.

2 Related work

The problem of multilingual polarity classification has already been addressed from different perspectives, such as monolingual sentiment analysis in a multilingual setting (Boiy and Moens, 2009), cross-lingual sentiment analysis (Brooke et al., 2009) or multilingual sentiment analysis (Balahur and Turchi, 2014). Banea et al. (2010) shows that including multilingual information can improve by almost 5% the performance of subjectivity classification in English. Davies and Ghahramani (2011) propose a language-independent model for sentiment analysis of Twitter messages, only relying on emoticons; that outperformed a *bag-of-words* Naive Bayes approach.

Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (WASSA 2015), pages 2–8,
Lisboa, Portugal, 17 September, 2015.

Cui et al. (2011) consider that not only emoticons, but also character and punctuation repetitions are language-independent emotion tokens. A different way of evaluating multilingual SA systems is posed by Balahur et al. (2014). They translate the English SemEval 2013 corpus (Nakov et al., 2013) into Spanish, Italian, French and German by means of machine translation (MT) systems. The resulting datasets were revised by non-native and native speakers independently, finding that the use of machine translated data achieves similar results as the use of native-speaker translations.

3 Multilingual sentiment analysis

Our goal is to compare the performance of supervised models based on *bag-of-words*, often used in SA tasks. We trained our classifiers using a L2-regularised logistic regression (Fan et al., 2008).

3.1 Feature Extraction

We apply Natural Language Processing (NLP) techniques for extracting linguistic features, using their total occurrence as the weighting factor (Vilares et al., 2014). 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[1] for 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 them.
- *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 1 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%[2] and 96.03% respectively. The multilingual tagger obtained 98.00% and 95.88% over the monolingual test sets.

These atomic sets of features can be combined to obtain a rich linguistic model that improves performance (Section 4).

3.2 Contextual features

Syntactic features

Dependency parsing is defined as the process of obtaining a dependency tree given a sentence. Let $S = [s_1 s_2 ... s_{n-1} s_n]$ be a sentence[3] of length *n*, where s_i indicates the token at the i^{th} position; a *dependency tree* is a graph of binary relations, $G = \{(s_j, m_{jk}, s_k)\}$, where s_j and s_k are the *head* and *dependent* tokens, and m_{jk} represents the syntactic relation between them. To obtain such trees, we trained an *en*, *es* and an *en-es* parser (Vilares et al., 2015b) using MaltParser (Nivre et al., 2007). In order to obtain competitive results for a 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 Labeled Attachment Score (LAS) of the Spanish and English monolingual parsers was 80.54% and 88.35%, respectively. The multilingual model achieved a LAS of 78.78% and 88.65% (significant improvement with respect to the monolingual model, using Bikel's randomised parsing evaluation comparator and $p < 0,05$). Figure 1 shows an example 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 a generalisation function, $g : W \rightarrow \{W, L, P, T\}$, a *generalised triplet* is defined as $(g(s_j), m_{ij}, g(s_k))$.

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

[2] 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).

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

	El	Cafe	Colombiano	taking	over	Newcastle	with	its	three	best
es	DET	NOUN	ADJ	X	X	X	X	X	X	X
en	NOUN	NOUN	NOUN	VERB	PTR	NOUN	ADP	PRON	NUM	ADJ
es-en	DET	NOUN	ADJ	VERB	ADP	NOUN	ADP	PRON	NUM	ADJ

Table 1: Performance of taggers on a code-switching sentence from Twitter: *adjective* (ADJ), *prepositions* and *postpositions* (ADP), *determinant* (DET), *noun* (NOUN), *particles* (PTR) *pronoun* (PRON), *verb* (VERB) and *other category* (X)

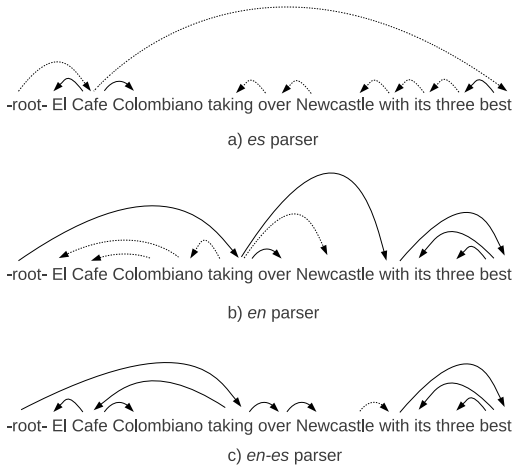

Figure 1: Example with the *en*, *es* and *en-es* dependency parsers. Dotted lines represent incorrectly-parsed dependencies

N-gram features

N-gram features capture shallow structure of sentences, identifying local relations between words (e.g. *'not good'* becomes *'not_good'*).

4 Experimental framework

The proposed sets of features and models are evaluated on standard monolingual corpora, taking accuracy as the reference metric. These 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.

4.1 Monolingual corpora

Two corpora are used to compare the performance of monolingual and multilingual models:

- *SemEval 2014 task B* corpus (Rosenthal et al., 2014): A set of English tweets[4] split into a training (8,200 tweets), development (1,416) and a test set[5] (5,752). Each tweet was manually classified as *positive*, *none* or *negative*.
- *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.

4.2 Multilingual corpora

These two test sets were merged to create a synthetic multilingual corpus. The aim was to compare the multilingual and the monolingual approach with language detection under this configuration. The unbalanced sizes of the test sets result in a higher performance when correctly classifying the majority language. We do not consider that 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.

4.3 Code-switching corpus

We created a polarity corpus with code-switching tweets based on the training collection[6] (*en-es*) presented by Solorio et al. (2014). Each word in the corpus is labelled with its language, serving as the starting point to obtain a collection of multilingual tweets. We first filtered the tweets containing both Spanish and English words, obtaining 3,062 tweets. Those were manually labelled by three annotators according to the SentiStrength strategy, a

[4]Due to Twitter restrictions some of the tweets are not available anymore, so the corpus statistics may vary slightly from those of other researchers that used the corpus.

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

[6]The test set was not released for the research community.

dual score *(p,n)* from 1 to 5 where *p* and *n* indicate the positive and the negative sentiment (Thelwall et al., 2010). Krippendorf's alpha coefficient indicated an inter-annotator agreement from 0.629 to 0.664 for negative sentiment and 0.500 to 0.693 for positive sentiment. To obtain the final score, we applied an average strategy with regular round: if $p > n$ then the tweet is labelled as *positive*, if $p < n$ then it is labelled as *negative* and otherwise it is labelled as *none*. After the transformation to the trinary scheme, we obtained a corpus where: the *positive* class represents the 31.45% of the corpus, the *negative* one represents a 25.67% and the remaining 42.88% belongs to the *none* class.

To the best of our knowledge, this is the first code-switching corpus with sentiment annotations.[7], which presents several challenges. It is an especially noisy corpus, were many grammatical errors occur in each tweet. There is also an overuse of subjective clauses and abbreviations (e.g. *'lol'*, *'lmao'*, ...) whose subjectivity was considered a controversial issue by the annotators. Finally, a predominant use of English was detected (`lang.py` classified 59.29% of the tweets as English). We believe this is because the Solorio et al. (2014) corpus was collected by downloading tweets for people from Texas and California.

5 Experimental results

5.1 Results on the English corpus

Features	en	pipe	en-es
Words (W)	**66.72**	66.71	66.22
Lemmas (L)	**66.74**	66.71	66.48
Psychometric (P)	62.52	**62.53**	61.47
PoS-tags (T)	51.82	51.80	**52.03**
Bigrams of W	60.99	61.00	**61.47**
Bigrams of L	61.75	**61.77**	61.32
Bigrams of P	**61.32**	**61.32**	60.41
Triplets of W	56.40	56.38	**57.84**
Triplets of L	58.69	58.67	**59.16**
Triplets of P	**58.26**	58.24	57.60
Combined (W,P,T)	68.52	**68.58**	68.48
Combined (L,P,T)	**68.43**	68.38	68.34
Combined (W,P)	68.72	**68.74**	68.52
Combined (L,P)	**68.57**	68.53	68.32

Table 2: Accuracy (%) on the SemEval 2014

Table 2 shows the performance of the three models on the SemEval test set. The differences between the monolingual model and the monolingual pipeline with language detection are tiny.

[7]Freely available in `grupolys.org/software/CS-CORPORA/cs-en-es-corpus-wassa2015.txt`

Features	1K test set			General test set		
	es	pipe	en-es	es	pipe	en-es
Words (W)	**56.60**	56.50	54.60	64.39	64.35	**64.59**
Lemmas (L)	56.40	56.30	**56.60**	64.45	64.48	**64.57**
Psychometric (P)	**54.70**	**54.70**	53.10	58.77	58.69	**59.50**
PoS-tags (T)	**48.90**	48.80	41.70	49.44	**49.49**	47.72
Bigrams of W	**52.90**	52.70	52.10	58.37	58.41	**58.66**
Bigrams of L	**54.00**	53.90	52.20	58.73	58.74	**59.29**
Bigrams of P	46.00	46.00	**47.00**	51.30	51.26	**53.22**
Triplets of W	**52.40**	52.20	44.60	54.26	54.41	**54.96**
Triplets of L	**54.40**	54.40	46.30	56.06	56.09	**56.38**
Triplets of P	**45.80**	**45.80**	47.50	50.00	49.44	**52.34**
Combined (W,P,T)	**60.00**	59.90	59.10	**66.43**	66.34	66.34
Combined (L,P,T)	**61.40**	**61.40**	59.20	**66.18**	66.10	66.12
Combined (W,P)	59.10	59.20	**59.60**	66.27	66.18	**66.28**
Combined (L,P)	59.80	**59.90**	59.30	**65.95**	65.89	65.92

Table 3: Accuracy (%) on the TASS test sets

This is due to the high performance of `lang.py` on this corpus, where only 6 tweets were misclassified as Spanish tweets. Despite of this issue, the *en-es* classifier performs very competitively on the English monolingual test sets, and the differences with respect to the *en* model range from 0.2 to 1.05 percentage points. With certain sets of features, consisting of triplets, the multilingual model even outperforms both monolingual models, reinforcing the validity of this approach.

5.2 Results on the Spanish corpus

With respect to the evaluation on the TASS 2014 corpus, the tendency seems to remain on the TASS 2014-1k, as illustrated in Table 3. It general terms the *es* model obtains the best results, followed by the *pipe* and the *en-es* models. In this version of the corpus, the 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 suggests that multilingual models present advantages when messages in different languages need to be analysed.

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

Features	pipe	en-es	pipe	en-es
Words (W)	**64.93**	64.20	64.55	**64.71**
Lemmas (L)	**65.03**	64.76	64.66	**64.72**
Psychometric (P)	**61.17**	60.02	59.03	**59.66**
PoS-tags (T)	**51.28**	50.23	**49.69**	48.11
Bigrams of W	59.55	**59.84**	58.63	**58.90**
Bigrams of L	**60.40**	59.73	59.00	**59.46**
Bigrams of P	**58.65**	58.08	52.19	**53.88**
Triplets of W	**55.65**	55.54	54.57	**55.21**
Triplets of L	**57.93**	56.92	56.31	**56.62**
Triplets of P	**56.08**	55.84	50.25	**52.81**
Combined (W,P,T)	**67.07**	66.85	**66.52**	**66.52**
Combined (L,P,T)	**67.17**	66.75	66.28	**66.30**
Combined (W,P)	**67.08**	66.97	66.39	**66.47**
Combined (L,P)	**67.03**	66.75	66.11	**66.12**

Table 4: Accuracy (%) on the multilingual test set

Features	en	es	pipe	en-es
Words (W)	**55.65**	47.65	52.74	54.87
Lemmas (L)	55.68	48.66	53.00	**56.37**
Psychometric (P)	53.04	43.63	50.69	**53.69**
PoS-tags (T)	**45.07**	39.32	44.71	43.17
Bigrams of W	54.31	47.45	51.67	**54.34**
Bigrams of L	55.03	48.92	52.16	53.63
Bigrams of P	**49.48**	40.46	46.08	46.86
Triplets of W	**52.55**	36.54	45.95	50.72
Triplets of L	**52.97**	44.68	48.99	50.42
Triplets of P	**48.14**	40.59	45.72	45.98
Combined (W,P,T)	**59.18**	48.27	56.53	58.52
Combined (L,P,T)	58.55	49.67	56.07	**59.11**
Combined (W,P)	58.72	49.90	56.40	**58.82**
Combined (L,P)	58.85	50.82	56.07	**59.34**

Table 5: Accuracy (%) on the code-switching set

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

5.3 Results on a synthetic multilingual corpus

Table 4 shows the performance both of the multilingual approach and the monolingual pipeline with language detection when analysing texts in different languages. On the one hand, the results show that using a multilingual model is the best option when Spanish is the majority language, probably due to a high presence of English words in Spanish tweets. On the other hand, combining monolingual models with language detection is the best-performing approach when English is the majority language. The English corpus contains only a few Spanish terms, suggesting that the advantages of having a multilingual model cannot be exploited under this configuration.

5.4 Results on the code-switching corpus

Table 5 shows the performance of the three proposed approaches on the code-switching test set. The accuracy obtained by the proposed models on this corpus is lower than on the monolingual corpora. This suggests that analysing subjectivity on tweets with code-switching presents additional challenges. The best performance (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 the ones obtained on the monolingual corpora, indicates 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, that made it possible to extract the sentiment of the texts in many of the cases.

6 Conclusions

We 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. The proposed approaches were: (a) a purely monolingual model, (b) a simple pipeline which used language identification techniques to determine the language of unseen texts (c) a multilingual model trained on a corpus that joins the two monolingual corpora. Experimental results reinforces the robustness of the multilingual approach under the three configurations.

Acknowledgments

This research is supported by the Ministerio de Economía y Competitividad (FFI2014-51978-C2) and Xunta de Galicia (R2014/034). The first author is funded by the Ministerio de Educación, Cultura y Deporte (FPU13/01180).

References

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

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

C. Banea, R. Mihalcea, and J. 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.

E. Boiy and M. 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 the International Conference RANLP-2009*, pages 50–54, Borovets, Bulgaria. ACL.

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

E. Cambria, B. Schuller, B. Liu, H. Wang, and C. Havasi. 2013b. Knowledge-based approaches to concept-level sentiment analysis. *IEEE Intelligent Systems*, (2):12–14.

A. Cui, M. Zhang, Y. Liu, and S. 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.

A. Davies and Z. Ghahramani. 2011. Language-independent Bayesian sentiment mining of {Twitter}. In *The 5th SNA-KDD Workshop'11 (SNA-KDD'11)*, San Diego, CA, August. ACM.

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

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

M. Lui and T. Baldwin. 2012. langid.py: An off-the-shelf language identification tool. In *Proceedings of the ACL 2012 system demonstrations*, pages 25–30. Association for Computational Linguistics.

R. McDonald, J. Nivre, Y. Quirmbach-Brundage, Y. Goldberg, D. Das, K. Ganchev, K. Hall, S. Petrov, Hao 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.

P. Nakov, S. Rosenthal, Z. Kozareva, V. Stoyanov, A. Ritter, and T. Wilson. 2013. SemEval-2013 Task 2: Sentiment Analysis in Twitter. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 312–320, Atlanta, Georgia, June. ACL.

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.

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

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

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

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

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

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, G. Paltoglou, D. Cai, and A. Kappas. 2010. Sentiment Strength Detection in Short Informal Text. *Journal of the American Society for Information Science and Technology*, 61(12):2544–2558, December.

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.

D. Vilares, M. Hermo, M. A. Alonso, C. Gómez-Rodríguez, and Y. Doval. 2014. LyS : Porting a Twitter Sentiment Analysis Approach from Spanish to English. In *Proceedings og The 8th International-Workshop on Semantic Evaluation (SemEval 2014)*, pages 411–415.

D. Vilares, M. A Alonso, and C. Gómez-Rodríguez. 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.

D. Vilares, M. A. Alonso, and C. Gómez-Rodríguez. 2015b. One model, two languages: training bilingual parsers with harmonized treebanks. *arXiv*, 1507.08449 [cs.CL].