Procesamiento del Lenguaje Natural, Revista nº 50 marzo de 2013, pp 93-100 recibido 17-12-12 revisado 07-02-13 aceptado 19-02-13

Lexicon-Based Sentiment Analysis of Twitter Messages in Spanish

Análisis de Sentimiento basado en lexicones de mensajes de Twitter en español

Antonio Moreno-Ortiz, Chantal Pérez Hernández
Facultad de Filosofía y Letras
Universidad de Málaga
{amo, mph}@uma.es

Resumen: Los enfoques al análisis de sentimiento basados en lexicones difieren de los más usuales enfoques basados en aprendizaje de máquina en que se basan exclusivamente en recursos que almacenan la polaridad de las unidades léxicas, que podrán así ser identificadas en los textos y asignárseles una etiqueta de polaridad mediante la cual se realiza un cálculo que arroja una puntuación global del texto analizado. Estos sistemas han demostrado un rendimiento similar a los sistemas estadísticos, con la ventaja de no requerir un conjunto de datos de entrenamiento. Sin embargo, pueden no resultar ser óptimos cuando los textos de análisis son extremadamente cortos, tales como los generados en algunas redes sociales, como Twitter. En este trabajo llevamos a cabo tal evaluación de rendimiento con la herramienta Sentitext, un sistema de análisis de sentimiento del español.
Palabras clave: análisis de sentimiento basado en lexicones, analítica de texto, textos cortos, Twitter, evaluación de rendimiento.

Abstract: Lexicon-Based approaches to Sentiment Analysis (SA) differ from the more common machine-learning based approaches in that the former rely solely on previously generated lexical resources that store polarity information for lexical items, which are then identified in the texts, assigned a polarity tag, and finally weighed, to come up with an overall score for the text. Such SA systems have been proved to perform on par with supervised, statistical systems, with the added benefit of not requiring a training set. However, it remains to be seen whether such lexically-motivated systems can cope equally well with extremely short texts, as generated on social networking sites, such as Twitter. In this paper we perform such an evaluation using Sentitext, a lexicon-based SA tool for Spanish.
Keywords: lexicon-based sentiment analysis, text analytics, short texts, Twitter, performance evaluation.

1 Introduction[1]

1.1 Approaches to Sentiment Analysis

Within the field of sentiment analysis it has become a commonplace assertion that successful results depend to a large extent on developing systems that have been specifically developed for a particular subject domain. This view is no doubt determined by the methodological approach that most such systems employ, i.e., supervised, statistical machine learning techniques. Such approaches have indeed proven to be quite successful in the past (Pang and Lee, 2004; Pang and Lee, 2005). In fact, machine learning techniques, in any of their flavors, have proven extremely useful, not only in the field of sentiment analysis, but in most text mining and information retrieval applications, as well as a wide range of data-intensive computational tasks. However, their obvious disadvantage in terms of functionality is their limited applicability to subject domains other than the one they were designed for. Although interesting research has been done aimed at extending domain applicability (Aue and Gamon 2005), such efforts have shown limited success. An important variable for these approaches is the amount of labeled text available for training the classifier, although they perform well in terms of recall even with relatively small training sets (Andreevskaia and

[1] This work is funded by the Spanish Ministry of Science and Innovation. LingMotif Project FFI2011-25893.

Bergler, 2007).

On the other hand, a growing number of initiatives in the area have explored the possibilities of employing unsupervised lexicon-based approaches. These rely on dictionaries where lexical items have been assigned either *polarity* or a *valence*[2], which has been extracted either automatically from other dictionaries, or, more uncommonly, manually. The works by Hatzivassiloglou and McKewon (1997) and Turney (2002) are perhaps classical examples of such an approach. The most salient work in this category is Taboada et al. (2011), whose dictionaries were created manually and use an adaptation of Polanyi and Zaenen's (2006) concept of Contextual Valence Shifters to produce a system for measuring the semantic orientation of texts, which they call SO-CAL(culator). This is exactly the approach we used in our Sentitext system for Spanish (Moreno-Ortiz et al., 2010).

Combining both methods (machine learning and lexicon-based techniques) has been explored by Kennedy and Inkpen (2006), who also employed contextual valence shifters, although they limited their study to one particular subject domain (the traditional movie reviews), using a "traditional" sentiment lexicon (the General Inquirer), which resulted in the "term-counting" (in their own words) approach.The degree of success of knowledge-based approaches varies depending on a number of variables, of which the most relevant is no doubt the quality and coverage of the lexical resources employed, since the actual algorithms employed to weigh positive against negative segments are in fact quite simple.

Another important variable concerning sentiment analysis is the degree of accuracy that the system aims to achieve. Most work on the field has focused on the *Thumbs up or thumbs down* approach, i.e., coming up with a positive or negative rating. Turney's (2002) work, from which the name derives, is no doubt the most representative. A further step involves an attempt to compute not just a binary classification of documents, but a numerical rating on a scale. The rating inference problem was first posed by Pang and Lee (2005), and the approach is usually referred to as *seeing stars* in reference to this work.

1.2 Sentiment Analysis for Spanish

Work within the field of Sentiment Analysis for Spanish is, by far, scarcer than for English.

Cruz et al. (2008) developed a document classification system for Spanish similar to Turney (2002), i.e. unsupervised, though they also tested a supervised classifier that yielded better results. In both cases, they used a corpus of movie reviews taken from the Spanish Muchocine website. Boldrini et al. (2009) carried out a preliminary study in which they used machine learning techniques to mine opinions in blogs. They created a corpus for Spanish using their Emotiblog system, and discussed the difficulties they encountered while annotating it. Balahur et al. (2009) also presented a method of emotion classification for Spanish, this time using a database of culturally dependent emotion triggers.

Finally, Brooke et al. (2009) adapted a lexicon-based sentiment analysis system for English (Taboada et al., 2006, 2011) to Spanish by automatically translating the core lexicons and adapting other resources in various ways. They also provide an interesting evaluation that compares the performance of both the original (English) and translated (Spanish) systems using both machine learning methods (specifically, SVM) and their own lexicon-based semantic orientation calculation algorithm, the above mentioned SO-CAL. They found that their own weighting algorithm, which is based on the same premises as our system (see below), achieved better accuracy for both languages, but the accuracy for Spanish was well below that for English.

Our system, Sentitext (Moreno-Ortiz et al., 2010, 2011), is very similar to Brooke et al.'s in design: it is also lexicon-based and it makes use of a similar calculation method for semantic orientation. It differs in that the lexical knowledge has been acquired semi-automatically and then fully manually revised from the ground up over a long period of time, with a strong commitment to both coverage and quality. It makes no use of user-provided, explicit ratings that supervised systems typically rely on for the training process, and it produces an index of semantic orientation based on weighing positive against negative text segments, which is then transformed into a ten-

[2] Although the terms *polarity* and *valence* are sometimes used interchangeably in the literature, especially by those authors developing binary text classifiers, we restrict the usage of the former to non-graded, binary assignment, and the latter is used to refer to an *n*-point semantic orientation scale.

point scale and a five-star rating system.

2 Sentiment Analysis with Sentitext

Sentitext is a web-based, client-server application written in C++ (main code) and Python (server). The only third-party component in the system is Freeling (Atserias et al., 2006, Padró, 2011), a powerful, accurate, multi-language NLP suite of tools, which we use for basic morphosyntactic analysis. Currently, only one client application is available, developed in Adobe Flex, which takes an input text and returns the results of the analysis in several numerical and graphical ways, including visual representations of the text segments that were identified as sentiment-laden[3]. Lexical information is stored in a relational database (MySQL).

Being a linguistically-motivated sentiment analysis system, special attention is paid to the representation and management of the lexical resources. The underlying design principle is to isolate lexical knowledge from processing as much as possible, so that the processors can use the data directly from the database. The idea behind this design is that all lexical sources can be edited at any time by any member of the team, which is facilitated by a PHP interface specifically developed to this end (GDB). This kind of flexibility would not be possible with the monolithic design typical of proof-of-concept systems.

2.1 Lexical resources

Sentitext relies on three major sources: the individual words dictionary (*words*), the multiword expressions dictionary (*mwords*), and the context rules set (*crules*), which is our implementation of Contextual Valence Shifters. The individual words dictionary currently contains over 9,400 items, all of which are labeled for valence. The acquisition process for this dictionary was inspired by the bootstrapping method recurrently found in the literature (e.g., Riloff and Wiebe, 2003, Gamon and Aue, 2005). Lexical items in both dictionaries in our database were assigned one of the following valences: -2, -1, 0, 1, 2. A more detailed description of these resources can be found in (Moreno-Ortiz et al., 2010).

The most similar sentiment analysis system to ours (Taboada et al., 2011) uses a scale from –5 to 5, which makes sense for a number of graded sets of near synonyms such as those given as examples by the authors (p. 273). In our opinion, however, as more values are allowed, it becomes increasingly difficult to decide on a specific one while maintaining a reasonable degree of objectivity and agreement among different (human) acquirers, especially when there is no obvious graded set of related words, which is very often the case.

There are two ways in which the original valence of a word or phrase can be modified by the immediately surrounding context: the valence can change in degree (intensification or downtoning), or it may be inverted altogether. Negation is the simplest case of valence inversion. The idea of Contextual Valence Shifters (CVS) was first introduced by Polanyi and Zaenen (2006), and implemented for English by Andreevskaia and Bergler (2007) in their CLaC System, and by Taboada et al. (2011) in their Semantic Orientation CALculator (SO-CAL). To our knowledge, apart from Brooke et al.'s (2009) adaptation of the SO-CAL system, to the best of our knowledge, Sentitext is the only sentiment analysis system to implement CVS for Spanish natively. Our context rules account both for changes of degree and inversion, and are stored in a database table which is loaded dynamically at runtime.

2.2 Global Sentiment Value

Sentitext provides results as a number of metrics in the form of an XML file which is then used to generate the reports and graphical representations of the data. The crucial bit of information is the Global Sentiment Value (GSV), a numerical score (on a 0-10 scale) for the sentiment of the input text. Other data include the total number of words, total number of lexical words (i.e., content, non-grammatical words), number of neutral words, etc.

To arrive at the global value, a number of scores are computed beforehand, the most important of which is what we call Affect Intensity, which modulates the GSV to reflect the percentage of sentiment-conveying words the text contains.

Before we explain how this score is obtained, it is worth stressing the fact that we do not count words (whether positive, negative, or neutral), but *text segments* that correspond to lexical units (i.e., meaning units from a lexicological perspective).

[3] The application can be accessed and tested online at http://tecnolengua.uma.es/sentitext

As we mentioned before, items in our dictionaries are marked for valence with values in the range -2 to 2. Intensification context rules can add up to three marks, for maximum score of 5 (negative or positive) for any given segment.

The simplest way of computing a global value for sentiment would be to add negative values on the one hand and positive values on the other, and then establishing it by simple subtraction. However, as others have noted (e.g., Taboada et al. 2011), things are fairly more complicated than that. Our Affect Intensity measure is an attempt to capture the impact that different proportions of sentiment-carrying segments have in a text. We define Affect Intensity simply as the percentage of sentiment-carrying segments. Affect Intensity is not used directly in computing the global value for the text, however, an intermediate step consists of adjusting the upper and lower limits (initially -5 and 5). The Adjusted Limit equals the initial limit unless the Affect Intensity is greater than 25 (i.e., over 25% of the text's lexical items are sentiment-carrying. Obviously, using this figure is arbitrary, and has been arrived at simply by trial and error. The Adjusted Limit is obtained by dividing the Affect Intensity by 5 (since there are 5 possible negative and positive valence values).

A further variable needs some explaining. Our approach to computing the GSV is similar to Polanyi and Zaenen's (2006) original method, in which equal weight is given to positive and negative segments, but it differs in that we place more weight on extreme values. This is motivated by the fact that it is relatively uncommon to come across such values (e.g. "extremely wonderful"), so when they do appear, it is a clear marker of positive sentiment. Other implementations of Contextual Valence Shifters (Taboada et al. 2011) have put more weight only on negative segments when modified by valence shifters (up to 50% more weight), operating under the so-called "positive bias" assumption (Kennedy and Inkpen 2006), i.e., negative words and expressions appear more rarely than positive ones, and therefore have a stronger cognitive impact, which should be reflected in the final sentiment score.

In our implementation, equal weight is placed to positive and negative values. However, we do not simply assign more weight to both extremes of the scale (-5 and 5), we place more weight on each increasingly toward both ends of the scale.

The resulting method for obtaining the Global Sentiment Value for a text is defined as:

$$GSV = \frac{(\sum_{i=1}^{5} 2.5 \cdot i \cdot N_i + \sum_{i=1}^{5} 2.5 \cdot i \cdot P_i) \cdot uB}{5 \cdot (LS - NS)} \quad (1)$$

where N_i is the number of each of the negative valences found, and P_i is the equivalent for positive values. The sum of both sets is then multiplied by the Affect Intensity (uB). LS is the number of lexical segments and NS is the number of neutral ones. Although not expressed in the equation, the number of possible scale points (5) needs to be added to the resulting score, which, as mentioned before, is on a 0-10 scale.

3 Task description

The evaluation experiment described in this paper was performed as conceived for the TASS Workshop on Sentiment Analysis, a satellite event of the SEPLN 2012 Conference. See Villena-Román et al., (2013) for a detailed description of the tasks involved.

4 Analysis of results

Although it might seem obvious, it is worth stressing that lexicon-based systems rely heavily on the availability of a certain number of words on which to apply the weighing operations. As described in section 2.2 above, Sentitext basically computes its GSV index by weighing the number and valences of polarity words and phrases against the number lexical segments found in the text. Although it does include threshold control (the Affect Intensity index discussed in 2.2 above) for varying text lengths, such threshold was designed to be applied to larger texts, considering "short" the average length of a media article or blog entry.

However, Twitter, with its 140 character limit, involves a radically different concept of "short". The average number of lexical segments per tweet, i.e., individual words and identified multiword expressions, that we obtained in our analysis of the test set was 14.1, whereas the average number of polarity-conveying segments was 5.5. This is a very high ratio indeed, implying that social networking sites are commonly used for expressing sentiments and opinions. This is in accord with what many scholars have found when analyzing SNS content (e.g., Siemens,

2011). Sentitext's Affect Intensity, i.e., the control threshold, is established at 25%, which, in our experience, is rarely reached except for extremely short texts with a high emotional load. These data are summarized in Table 1.

	N	%	AVG/tweet
Lexical	857,727	100	14.1
Polarity	337,238	39,32	5.5
Neutral	520,489	60,68	8.6

Table 1: Polarity of text segments

It is therefore not surprising that our analysis of this Twitter test set throws an average Affect Intensity of 19.22, which is extremely high, especially if we bear in mind that 38.4% of the tweets have an Affect Intensity of 0, that is, they are neutral. As for the tweets classification task itself, we show and discuss the results in the following section, where we also offer figures of a more typical evaluation scenario in which texts are categorized as negative, neutral, or positive, i.e., there is no intensification for polarity categories and no distinction between the NEU and NONE categories (both are considered as neutral).

4.1 Three levels + NONE test

Table 2 below summarizes the hit rate for each of the categories, as well as overall.

	N	Hits	H %	Misses	M %
N	15,840	8,848	55.86	6,992	44,14
NEU	1,302	647	49.69	655	50.31
P	22,231	13,284	59.75	8,947	40.25
NONE	21,411	84	0.39	21,327	99.61
Total	60,784	21,327	37.61	37,921	62.39

Table 2: Hit rate for 3L+N test

Results are above average for polarity categories, but not so much for neutral and especially for the NONE category, with just a 0.39% hit ratio. The reason for this is that we decided to classify tweets as belonging to this category exclusively when they were essentially void of content, for example, those that contained just a URL. Clearly this is not what was meant, but we have to say that even after analyzing the correct results, the difference between NEU and NONE is still not clear.

The first conclusion that can be drawn, as far as actual performance is concerned, is that Sentitext has an excessive tendency to assign middle-of-the-scale ratings, both when the correct assignment is negative and positive. Since our tool does not classify, but simply assign a rating on a scale, the actual classification implied deciding on the scale boundaries for each of the categories. Table 3 below shows the boundaries we selected for this test.

Category	GSV Range
P+	GSV>8
P	GSV<=8
NEU	GSV<=5.4
N	GSV<=4
N+	GSV<=2
NONE	No content

Table 3: GSV ranges used for classification

An obvious way in which we could have optimized these ranges and obtained better results would have been contrasting the training set results with ours. This would have also softened the impact caused by the NEU-NONE issue.

Table 4 below offers performance results in terms of the usual metrics for classifiers.

	Precision	Recall	F
N	0.559	0.691	0.618
NEU	0.497	0.023	0.043
P	0.598	0.688	0.639
NONE	0.004	1	0.008

Table 4: Evaluation metrics for 3L+N test

As expected, these figures are extremely low for the NEU and NONE categories. The high recall rate for the NONE category is due to the fact that we only classified 84 tweets under this category, all of which were correct. Of course the harmonizing F-measure is very low anyway.

Even the metrics for the negative and positive are relatively low in comparison with previous tests (e.g., Moreno-Ortiz et al., 2011). We believe this may be due to the short length of the texts, and it is something will seek to improve in the future.

4.1.1 Unofficial 3L(-N) test

Since we do not think adding a NONE category serves any practical purpose, we decided to perform the same test removing the NONE category, in order to obtain more useful conclusions as to performance in a real-world scenario, and also to measure the precise impact

that the NONE issue had on the overall performance.

Table 5 below offers the hit rate for each category, which are obviously unchanged from the 3L+N test, but show important differences with the NEU one.

	N	Hits N	H %	Misses	M %
N	15,840	8,848	55.86	6,992	44.14
NEU	22,713	15,709	69.16	7,004	30.84
P	22,231	13,284	59.75	8,947	40.25
Total	60,784	37,841	62.25	22,943	37.75

Table 5: Hit rates for 3L(-N) test

By removing the NONE category, the overall hit rate rises from 37.61% to 62.25%, a difference of 24.64%, as a consequence of dramatic improvement of the hit rate for the NEU category.

In terms of precision and recall, we obtain a proportional improvement for the NEU category, as shown in Table 6 below.

	Precision	Recall	F
N	0.559	0.691	0.618
NEU	0.692	0.548	0.611
P	0.598	0.688	0.639

Table 6: Evaluation metrics for 3L(-N) test

4.2 Five levels + NONE test

Again, the influence of the NEU-NONE issue is strong, as can be clearly seen both in the figures in Figure 2, where it is also noticeable the better performance for negative cases than positive ones. Table 7 summarizes the hit rate for each of the categories and overall.

	N	Hits	H %	Misses	M %
N+	4,552	955	20.98	3,597	79.02
N	11,281	5,075	44.99	6,206	55.01
NEU	1,300	647	49.77	653	50.23
P	1,483	760	51.25	723	48.75
P+	20,741	2,643	12.74	18,098	87.26
NONE	21,409	84	0.39	21,325	99.61
Total	60,766	10,164	16.73	5,0602	83.27

Table 7: Hit rate for 5L+N test

The overall hit rate is in this case extremely low, nearly half as in the 3L+N test (16.73%). This is the result of the above-mentioned tendency that Sentitext displays toward middle-of-the-scale values, since most misses (apart from the one caused by NONE) come from classifying N+ as N and P+ as P. It is quite apparent that our current implementation of the GSV index needs a revision in order to make extreme values at both ends of the scale more easily attainable.

Of course, another way in which we could overcome this issue would be simply to use different ranges in our classification scale (see Table 4 above). This quick-and-dirty approach may be worth trying, at least as an interim solution whenever classification is required.

The standard evaluation metrics are provided in Table 8 below.

	Precision	Recall	F
N+	0.210	0.373	0.269
N	0.450	0.496	0.472
NEU	0.498	0.023	0.043
P	0.512	0.047	0.085
P+	0.127	0.885	0.223
NONE	0.004	1.000	0.008

Table 8: Evaluation metrics for 5L+N test

The figures clearly reflect the low performance, which falls below 50% in all cases, and especially at both extremes (N+ and P+). It is interesting, though, how recall for the P+ category is particularly high in relation to precision, even proportionally to that of N+. This is because our analyzer only assigned 2,643 cases to this category, which in fact had 21,409 cases, a surprising figure that contrasts with the 4,552 cases for N+. Table 9 below summarizes the official results and provides percentages for each category.

	N	%	3L%
N+	4,552	7.49	26.05
N	11,281	18.56	
NEU	1,300	2.14	
P	1,483	2.44	36.58
P+	20,741	34.13	
NONE	21,409	34.23	34.23
Total	60,766	100	100

Table 9: Official results for each category

Any number of conclusions can be drawn from these numbers.

4.2.1 Unofficial 5L(-N) test

As we did before, we show the hypothetical results that we would obtain if the NONE

category were to be removed. Tables 10 shows the hit rate in this scenario.

	N	Hits	H %	Misses	M %
N+	4,552	955	20.98	3,597	79.02
N	11,281	5,075	44.99	6,206	55.01
NEU	22,709	15,709	69.18	7,000	30.82
P	1,483	760	51.25	723	48.75
P+	20,741	2,643	12.74	18,098	87.26
Total	60,766	25,142	41.38	35,624	58.62

Table 10: Hit rate for 5L(-N)

And finally, precision and recall figures:

	Precision	Recall	F
N+	0.210	0.373	0.269
N	0.450	0.496	0.472
NEU	0.692	0.548	0.612
P	0.512	0.047	0.085
P+	0.127	0.885	0.223

Table 11: Evaluation metrics for 5L(-N)

As before, precision and recall are the same for all categories except NEU, which rises significantly in precision, and extremely in recall. Hit rate improves in the same proportion as in the 3L(-N) test.

5 Conclusions

Performing this test has been extremely useful to identify weaknesses in our current implementation of Sentitext's Global Sentiment Value. On the one hand, this test confirms our initial impressions after carrying out some informal tests with Twitter messages, that GSV is strongly affected by the number of lexical units available in the text (or the lack of them, rather). On the other hand, we have also confirmed Sentitext's tendency to assign middle-of-the-scale ratings, or at least avoid extreme values, which is reflected on its poor performance for the N+ and P+ classes, most of which were assigned to the more neutral N and P classes. This happens despite the fact that our GSV calculation places more weight on extreme values. Conversely, we found a relatively high proportion of polarized lexical segments found (high Affect Intensity). This is something that could not be inferred from the results of machine learning classifier. Even with a high proportion of neutral messages, these numbers clearly support the claims of many social media analysts that social networking sites, are used mainly to circulate news and express emotions about them. But our data also indicate that they tend to avoid strong language to convey their opinions, relying more on mild expression, implicature, and shared knowledge.

The third important conclusion is that differentiating between neutral and no polarity may not be the best decision, since it is not clear what the difference is. In fact, after checking the official assignment of these tags to the test set, it seems to us completely random. Therefore, it is very difficult to obtain good results in these two categories. Furthermore, there really is no need whatsoever to make this distinction from a practical perspective.

6 References

Andreevskaia, A., and Bergler, S. (2007). CLaC and CLaC-NB: knowledge-based and corpus-based approaches to sentiment tagging. *Proceedings of the 4th International Workshop on Semantic Evaluations* (pp. 117–120). Prague, Czech Republic: Association for Computational Linguistics.

Atserias, J., Casas, B., Cornelles, E., González, M., Padró, L., and Padró, M. (2006). FreeLing 1.3: Syntactic and semantic services in an open-source NLP library. *Proceedings of the fifth international conference on Language Resources and Evaluation*. Presented at the LREC 2006, Genoa, Italy: ELRA.

Aue, A., and Gamon, M. (2005). Customizing Sentiment Classifiers to New Domains: A Case Study. Presented at the Recent Advances in Natural Language Processing (RANLP), Borovets, Bulgaria.

Balahur, A., Kozareva, Z., and Montoyo, A. (2009). Determining the Polarity and Source of Opinions Expressed in Political Debates. *Proceedings of the 10th International Conference on Computational Linguistics and Intelligent Text Processing*, CICLing '09 (pp. 468–480). Berlin, Heidelberg: Springer-Verlag.

Boldrini, E., Balahur, A., Martínez-Barco, P., and Montoyo, A. (2009). EmotiBlog: an Annotation Scheme for Emotion Detection and Analysis in Non-traditional Textual Genres. *Proceedings of The 2009 International Conference on Data Mining*

(pp. 491–497). Presented at the DMIN 2009, Las Vegas, USA: CSREA Press.

Brooke, J., Tofiloski, M., and Taboada, M. (2009). Cross-Linguistic Sentiment Analysis: From English to Spanish. *Proceedings of RANLP 2009, Recent Advances in Natural Language Processing*. Presented at the RANLP 2009, Borovets, Bulgaria.

Cruz, F., Troyano, J. A., Enriquez, F., and Ortega, J. (2008). Clasificación de documentos basada en la opinión: experimentos con un corpus de críticas de cine en español. *Procesamiento del Lenguaje Natural*, (41), 73–80.

Gamon, M., and Aue, A. (2005). Automatic identification of sentiment vocabulary: exploiting low association with known sentiment terms (pp. 57–64). Ann Arbor, Michigan: Association for Computational Linguistics.

Hatzivassiloglou, V., and McKeown, K. R. (1997). Predicting the semantic orientation of adjectives. *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics* (pp. 174–181). Madrid, Spain: Association for Computational Linguistics.

Kennedy, A., and Inkpen, D. (2006). Sentiment classification of movie reviews using contextual valence shifters. *Computational Intelligence*, *22*(2), 110–125.

Moreno-Ortiz, A., Pineda Castillo, F., and Hidalgo García, R. (2010). Análisis de Valoraciones de Usuario de Hoteles con Sentitext: un sistema de análisis de sentimiento independiente del dominio. *Procesamiento de Lenguaje Natural*, *45*, 31–39.

Moreno-Ortiz, A., Pérez-Hernández, C., and Hidalgo-García, R. (2011). Domain-neutral, Linguistically-motivated Sentiment Analysis: a performance evaluation. *Actas del XXVII Congreso de la Sociedad Española para el Procesamiento del Lenguaje Natural* (pp. 847–856).

Padró, L. (2011). Analizadores Multilingües en FreeLing. *Linguamatica*, *3*(2), 13–20.

Pang, B., and Lee, L. (2004). A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts. *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics* (p. 271). Barcelona, Spain: Association for Computational Linguistics. Retrieved from http://portal.acm.org/citation.cfm?id=1218990anddl=GUIDEandcoll=GUIDEandCFID=80308782andCFTOKEN=73139236

Pang, B., and Lee, L. (2005). Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. *Proceedings of ACL 2005* (pp. 115–124). Presented at the ACL.

Polanyi, L., and Zaenen, A. (2006). Contextual Valence Shifters. *Computing Attitude and Affect in Text: Theory and Applications* (pp. 1–10). Dordrecht, The Netherlands: Springer.

Riloff, E., and Wiebe, J. (2003). Learning extraction patterns for subjective expressions. *Proceedings of the 2003 conference on Empirical methods in natural language processing*, EMNLP '03 (pp. 105–112). Stroudsburg, PA, USA: Association for Computational Linguistics.

Siemens, G. (2011, July 30). Losing interest in social media: there is no there there. *Elearnspace*. Retrieved from http://www.elearnspace.org/blog/2011/07/30/losing-interest-in-social-media-there-is-no-there-there/

Taboada, M., Brooks, J., Tofiloski, M., Voll, K., and Stede, M. (2011). Lexicon-Based Methods for Sentiment Analysis. *Computational Linguistics*, *37*(2), 267–307.

Turney, P. D. (2002). Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)* (pp. 417–424). Presented at the ACL 2002, Philadelphia, USA.

Villena-Román, J., García-Morera, J., Moreno-Garcia, C., Ferrer-Ureña, L., Lana-Serrano, S., González-Cristobal, J. C., Westerski, A., Martínez-Cámara, E., García-Cumbreras, M. A., Martín-Valdivia, M. T., Ureña-López L. A. 2012. TASS-Workshop on Sentiment Analysis at SEPLN. *Procesamiento de Lenguaje Natural*, 50.