Domain-neutral, Linguistically-motivated Sentiment Analysis: a performance evaluation

Evaluación de un sistema de análisis de sentimiento basado en conocimiento lingüístico e independiente del dominio

Antonio Moreno Ortiz, Chantal Pérez Hernández, Rodrigo Hidalgo García
Facultad de Filosofía y Letras
Universidad de Málaga
Campus de Teatinos
29071 Málaga
{amo,mph,rodrigo.hidalgo}@uma.es

Abstract: Within the field of sentiment analysis it has become commonplace the assertion that successful results depend to a large extent on developing systems specifically designed for a particular subject domain. In this paper we challenge this view by evaluating a domain-independent sentiment analysis system against a multiple-domain opinion corpus. The results show that high performance can be achieved by relying entirely on high quality, manually acquired, linguistic knowledge.

Keywords: sentiment analysis, opinion mining, multiple-domain opinion corpus.

Resumen: En el campo del análisis de sentimiento es común encontrar la afirmación de que para obtener buenos resultados es necesario emplear sistemas específicamente diseñados para un dominio temático en particular. En este trabajo ofrecemos una visión opuesta mediante la evaluación de un sistema de análisis de sentimiento independiente del dominio, que realizamos utilizando un corpus de opinión de múltiples dominios. Los resultados muestran que es posible obtener un alto rendimiento empleando exclusivamente recursos de conocimiento lingüístico de alta calidad obtenidos de forma manual.

Palabras clave: análisis de sentimiento, minería de opinión, corpus de opinión multi-dominio.

1 *Introduction*

Within the field of sentiment analysis it has become commonplace the 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 & Lee, 2004; Pang & Lee, 2005; Aue & Gamon, 2005).

Machine learning algorithms, in any of their flavors, have indeed 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 & 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 & Bergler, 2007).

On the other hand, a growing number of initiatives in the area have explored the possibilities of employing unsupervised knowledge-based approaches. These rely on a dictionary where lexical items have been assigned a valence, either extracted automatically from other dictionaries, or, more uncommonly, manually acquired. The works by Hatzivassiloglou & McKewon (1997) and Turney (2002) are perhaps classical examples of such an approach.

Hybrid, i.e., semi-supervised, approaches have also been employed, as in Goldberg & Zhu (2006), where both labeled and unlabeled data are used.

Extraction of lexical cues for semantic orientation (i.e., polarity) is usually performed semi-automatically, for example by Mutual Information scores obtained from adjectives or adverbs, which are the most obvious word classes to convey subjective meaning. To a lesser extent, nouns (e.g. Riloff et al., 2003) and verbs (e.g. Riloff & Wiebe, 2003) have also been used to identify semantic orientation.

The degree of success of such approaches varies depending on a number of variables, of which the most salient is no doubt the quality and coverage of the lexical resources employed, since the actual algorithms employed to weigh positive against negative segments is in fact quite simple.

To summarize, unsupervised, statistics-based approaches tend to be of limited application and tend to achieve good recall, but low precision, whereas unsupervised, knowledge-based approaches display the opposite results: they are good at precision but may miss many sentiment-laden text segments (Andreevskaia & Bergler, 2007).

1.1 Thumbs vs. stars

Another important variable concerning Sentiment Analysis is the degree of accuracy that the system attains to achieve. Most work on the field has focused on the *Thumbs up or thumbs down* approach, i.e., coming up with positive or negative rating. Turney's (2002) work is no doubt the most representative.

A further step consists of computing not just a binary classification of documents, but a numerical rating. The *rating inference problem* was first posed by Pang & Lee (2005), and the approach is usually referred to as "seeing stars" in reference to this work, where they compared different variants of the original SVM binary classification scheme aimed at supporting *n*-ary classification. Shimada & Endo (2008) and Gupta et al. (2010) further elaborated on the multi-scale issue by tackling *multi-aspect*, i.e., pinpointing the evaluation of multiple aspects of the object being reviewed, a feature we regard as essential for high-quality, fine-grained sentiment analysis, but one that requires very precise topic identification capabilities.

Our system, Sentitext (Moreno-Ortiz et al. 2010a), is knowledge-based, and its knowledge sources (the individual words lexicon, phrases lexicon and context rules set) have all been manually acquired, using both dictionaries and corpora. 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 five-point scale.

Rating systems, in any of their forms, are used by a large number of web sites, including some of the largest ones: eBay, Amazon, Netflix, IMDb, and many other user reviews sites all use multi-point scale rating systems, 5-point scales being the most common in the Internet.

Users' familiarity with this 5-star rating system is the main reason why we decided to implement it in Sentitext. Other than that, it does present a number of difficulties when it comes to weighing results against human ratings. We describe these in section 4.1 below.

1.2 Sentiment Analysis for Spanish

Our initial evaluation of Sentitext (Moreno-Ortiz et al., 2010b) only goes to add more evidence to our claim that knowledge-based, linguistically-motivated sentiment analysis achieves good precision results. This evaluation, however, focused on hotel reviews specifically. In the present paper we address the issue of multiple domains by analyzing a wider range of topics.

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 *Muchocine* web site, which is available for free use, thus focusing on a particular domain. 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.

It is quite apparent that advances within the field of Sentiment Analysis for Spanish are, by far, more scarce than studies carried out for English. Besides, most studies focus on specific domains, typically movie reviews.

2 *Dealing with ambiguous meaning*

There is little doubt that lexical and structural ambiguity has repeatedly proven to be the Achilles' heel of many Natural Language Processing applications. A wide range of solutions have been proposed, whether based on —mainly lexically-motivated— linguistic theories and approaches, or ad-hoc computational methods.

Sentiment Analysis has not escaped this ubiquitous fly in the ointment. The affective polarity of many words can shift to neutral, or even be inverted altogether, depending on the context. In spoken language, a mere change of intonation or other prosodic features, even very slightly, can be an indication that lexical meaning should not be taken literally, or perhaps that irony or sarcasm is being conveyed, or in fact any of a wide range of human emotions.

Fortunately, we do not have to deal with such prosodic subtleties in written text, but this does not mean that texts do not present formidable challenges for automatic analysis, quite the contrary.

2.1 Subjectivity and Sentiment Analysis

The subjectivity/objectivity axis, in particular, is a well-known source of difficulty, having received attention in the literature, Akkaya & Wiebe (2009). As Kim & Hovy (2004) point out, differentiating fact from opinion is not easy even for humans: personal comments and points of view can be disguised as fact in a number of ways in written discourse.

In our view, this axis is not so relevant to sentiment analysis as it is to opinion mining specifically. Words like *disease* or *headache* have a negative polarity whether they are used subjectively, as in the examples in (1), or objectively, as in (2) below:

(1) He is a **disease** to every team he has gone to.
Converting to SMF is a **headache**.

(2) Early symptoms of the **disease** include severe **headaches**.[1]

The point we wish to make is that the result in terms of polarity classification of a user review, will not be different whether, objectively or subjectively, certain features of the entity being discussed are deemed not adequate; what is relevant is the fact that the user raised that particular fact/opinion about the entity.

2.2 Lexical ambiguity

Polarity, however, may or may not be present depending on the word sense selected in a particular context, which directly affects the result in sentiment analysis. In (3) below the word *breeze* has neutral polarity, whereas in (4) the word is positive, meaning "easy to perform".

(3) The **breeze** caressed her hair.

(4) Getting the work done was a **breeze**.

A straightforward approach to word-sense disambiguation for sentiment analysis tasks is sense labeling. Such approach has been used in the past with various degrees of success. The most salient work in this respect is that of Esuli and Sebastiani (2006), who have developed SentiWordnet, a version of WordNet specially designed for sentiment analysis.

Our system lacks such labeling, and the word lexicon it employs contains no other information apart from the valence. This monosemous approach is apparently extremely simplistic, but we have found that a large part of ambiguous cases can be resolved by using the two other main lexical knowledge sources that Sentitext uses: the phrase lexicon and the valence shifters set, described in section 3 below.

2.3 Domain-related ambiguity

Performing tests on multiple domains has allowed us to find out some interesting observations, among which we would like to point out a distinction in nature of two different types of reviews according to the object being assessed:

1. Type 1: those that relate to something with conceptual content: movies, books, music, etc.

[1] The examples are taken from Akkaya & Wiebe (2009).

2. Type 2: those that do not: consumer goods and artifacts in general.

In a movie review, for example, the reviewer is likely to discuss and evaluate any number of entities, such as the movie itself, the actors, director, producer, and so on, which directly contribute to the overall opinion on the movie, but also they will comment on the entities and events occurring within the movie itself, which will be described as positive or negative depending on their specific nature, rather than the reviewer's view. Thus a review about a movie or novel depicting tragic events is bound to contain more negative language than a review about a romantic comedy. Reviewers will sometimes distinguish their plot summary from the review itself, but more commonly they will seamlessly integrate both. As a result, highly sophisticated entity recognition and topic identification techniques are needed to differentiate and isolate each type of content.

This is not to say that evaluative language irrelevant to the object being discussed may not appear in Type 2 reviews. A user reviewing a certain appliance may relate any number events of their own life, but this type of content is certainly much more limited.

3 *Contextual Valence Shifters*

Simply accounting for negative and positive words and phrases found in a text will not be enough. There are two ways in which their valence 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 & Zaenen (2006), and implemented for English by Andreevskaia & Bergler (2007) in their CLaC System. To our knowledge, Sentitext is the first and the only sentiment analysis system to implement CVS for Spanish.

Our CVS system is implemented in what we call *Context Rules*, which are expressed as the following data structure:

- Unit Form: Freeling-compliant morpho-syntactic definition of the item being modified (e.g.: "AQ").
- Unit Sign: polarity of the item being modified (e.g. "+").
- CVS Definition: modifier definition (e.g.: "muy").
- CVS Position: position of the modifier (e.g. "L" for *left*).
- CVS Span: maximum number of words where the modifier can be found from the modified item.
- Result: valence result of the modification: INV (valence inversion), INT*n* (valence intensification of *n*), or DOW*n* (valence downtoning of *n*).

This system allows us to describe fairly elaborated context rules, for instance having multiword modifiers such as "no tener nada de + AQ" or "el peor + NC + del mundo".

4 *Evaluation design*

Our methodology consisted basically of the following steps. After sample selection, all texts in the sample were rated independently by three human users, using a 5-star-rating scale. Users were provided with the review text only. Depending on the review source, sometimes the review includes a title, which can be very representative of semantic orientation, and sometimes a label, or self-rating, in whichever form (star system or numeric value). Such elements, when available, were removed from the texts, as they can clearly prime users in a certain direction.

Next, we calculated the average value of these three human-generated ratings and, finally, we compared the average to Sentitext's rating.

4.1 Sample reviews

For our experiment, we selected a sample[2] consisting of four sets of 25 reviews each, taken from our opinion corpus (*COE: Corpus de Opinión del Español*). Each set belongs to a different domain: (a) movie reviews, (b) books and music reviews, (c) consumer goods reviews, and (d) electronic products reviews. All the selected review texts were roughly equal in size, originally written in Spanish, and classified in one of these four sets according to the topic they dealt with to ensure domain homogeneity. Table 1 shows the different sources for each domain; since we would not be using self-assigned ratings (i.e., users' labels), we were able to use a wide variety of sources.

[2] The sample reviews used in this paper are available for download in XML format from our server at http:/tecnolengua.uma.es/coe/xml/sepln2011sample.tar.gz.

Domain	Sources
Movies	http://www.muchocine.net http://www.labutaca.net http://www.precriticas.com
Books and music	http://www.criticadelibros.com http://www.elcultural.es http://www.ciao.es
Consumer goods	http://www.ciao.es http://www.doyoo.es
Electronics	http://www.ciao.es http://www.videojuegos.tv http://www.pcactual.com http://www.quo.es http://www.homotecnologicus.com http://kindlespain.es

Table 1: Sample reviews domains and sources

4.2 Computing ratings

Sentitext computes the polarity rating of a text by weighing the valences it assigns to positive and negative text segments that it has been able to identify. The valence of lexical items, both individual words and phrases, is obtained from the manually-assigned valence they have in the lexicons, and can later be modified by the context rules as explained in section 3 above. Stored valences for lexical items range from -3 to 3, which can be intensified by context rules for a maximum of -5, for negative segments, and 5 for positive ones.

A global value (*gValue*) is returned based on a 0-10 scale, and then this *gValue* is converted to a 5-point scale. The current interface returns both results.

This *gValue* is calculated as the sum of all identified affect-loaded text segments, extreme values having a heavier weight than middle ones. Affect intensity, i.e. the proportion of affect-loaded segments to the overall number of lexical words, is also factored in. We do this in order to prevent mainly neutral text to be awarded a high semantic orientation. Weights and affect intensity values have been chosen arbitrarily, and optimized by trial-and-error.

Conversion of this *gValue* to a 5-star rating system is not as straightforward as it might seem, due to the quirks that such systems are known to have. An odd number of points might suggest that the middle point (3 in a 5-point scale) would indicate neutrality in the reviewer's opinion; however, this is not the case. 3 stars is used to rate a product or service as "Good" or "OK", as is sometimes explicitly found on review websites. This means that there are 2 possible negative ratings versus 3 available positive ratings, which involves a certain bias toward positive responses. It also means that distances between points must be interpreted differently: the relative distance between 3 and 5 stars is "shorter" than that between 1 and 3: the former does not change polarity (both ratings are positive), whereas the latter marks a clear change in polarity.

As a result, assessing precision of an automated rating inference system, such as the one we put to the test in this paper, clearly needs to take this into account. Table 2 below specifies the terms under which we define precision, where the above remarks have been factored in.

(i)	d(*xy*) < 1	HIT
(ii)	d(*xy*) < 2	NEAR HIT
(iii)	d(*xy*) = 2 AND (*x* =3; y=5) OR (x=5; *y* = 3)	NEAR HIT
(iv)	ELSE	MISS

Table 2: Precision values for assessing 5-star rating agreement

Thus, we rank agreement on the distance[3] *d* between any two star ratings *x* and *y* as falling into one of these four cases. Rank (i), a full hit, is defined as a distance of less than one star. A near hit is obtained as either obtaining a distance shorter than 2 (ii) or greater than 2 when the lower value is greater than 2 (iii). The rest of the cases are deemed a miss (iv). For 2-star distances, the challenging case, this scheme results in the following:

- Rating1—1star; Rating2—3star: MISS.
- Rating1—2star; Rating2—4star: MISS.
- Rating1—3star; Rating2—5star: NEAR HIT.

5 *Results*

5.1 Human-human agreement

Agreement between our three human raters was generalized: in 48% of the texts, they agreed completely (the three of them assigned the same number of stars to the reviews) and in 47 cases, two raters agreed whereas the other one did not, but the difference was only one star. A difference of 2 stars took place in only 5% of

[3] We do not use half-star notation. However, averages will produce non-integer values, hence the notation used in Table 2.

the texts, where the three ratings given by our analysts were different. These five cases of “human disagreement” deserve a closer look, due to the non-symmetrical nature of the star-rating system we already mentioned in the previous section (4.2). For two of the texts, the ratings given ranged from 3 to 5 stars, which does not involve a change in: the three ratings are positive. However, in the other three cases, the difference ranged from 1 to 3, meaning that one of the raters found the text *very* negative, another one rated it as *fairly* negative and the third one simply as positive or good.

These figures mean that, with the exception of three of the texts, the three human raters agreed as far as the polarity of the texts was concerned; they mainly differed with regard to the perception of the *intensity* of the emotion portrayed in the texts (fairly positive vs. strongly positive, for instance).

There seems to be, however, some domain-dependency for the degree of agreement: it is slightly higher in the last two sets of texts (consumer goods and electronic products reviews) than in the first two groups (books & music and movies reviews). Furthermore, the only five cases in which the three human raters differed among them belong to these two domains. This may be explained by the fact that these types of reviews (books, music or cinema) tend to be more subjective in nature, what makes the texts harder to rate even for human analysts, although the overall polarity of the text is not really affected.[4]

5.2 Human-machine agreement

Having a quick glance at the results yielded by Sentitext immediately shows us a high degree of agreement between our human raters and the automatic analysis carried out, in line with Moreno-Ortiz et. al (2010b), with a hit rate of 90%. It is, however, worthwhile to study in detail those cases in which human ratings and Sentitext did not agree.

In those texts where human raters and Sentitext differed by 2 points or more with a change in the polarity assigned to the text, the majority of cases are texts to which Sentitext assigns a higher rating than human raters: typically humans would assign 1 star (very negative) to a text which Sentitext rates as having 3 stars, thus being “good”. (see Table 3).

Reviewers’ Average → Sentitext Hit/Near Hit	90%
Reviewers’ Average → Sentitext Missed	10%
Missed: Reviewers (-) → Sentitext (+)	7%
Missed: Reviewers (+) → Sentitext (-)	3%

Table 3: Reviewers’ Average and Sentitext agreement.

6 Discussion of results

It seems apparent that Sentitext is giving a higher value to texts which reviewers are sure to mark as negative, thus we assume that Sentitext is finding positive segments which, for the purposes of the product’s review as such, seem pointless to our raters. This is especially true of texts that are longer than average and written in a narrative style, either reviewing previous works/products which used to be good just to end up claiming, in a short paragraph, that the new model is inadequate, or in cases where the author tells us about the successful career of a filmmaker who happens to have released, this time, a bad quality movie. In other words, a large amount of positive segments are not used to evaluate the product concerned, thus being ignored by our reviewers but not by Sentitext. Giving higher prominence to parts of the texts that, from the discursive point of view, possess more evaluative potential, i.e. the beginning and end of the text, could be a possible solution to this automatic evaluation inaccuracy, as Taboada & Grieve (2004) suggest. We need to consider that discursive elements alone can determine an opinion, for instance a negative one, even when no overtly-marked negative words are used, for example:

(5) Si quieres ver una película, mejor pasar a otra cosa, si tienes curiosidad por el trabajo que ha realizado este colectivo, adelante.

Apart from these elements of discourse that may affect the way Sentitext rates texts, we have encountered other evidence, from the lexical point of view, to explain human-Sentitext mismatch. Sometimes, words that are

[4] In fact, some opinion-tracking systems rate subjectivity and sentiment separately (Godbole et al., 2007).

clearly positive are used to give negative evaluation and vice versa. Consider the following examples extracted from our sample texts:

(6) …se presta gustosamente a hacer el ridículo.

(7) …maltrata propuestas más valiosas para abrazar suplicios como este.

(8) …todo una ganga teniendo en cuenta la magnífica calidad del producto; …este es el regalo perfecto....

It is clear that the use of positive words in (6) and (7) does not counterbalance the negative impact of the judgment as a whole. Example (8) is different since, out of context, there is no reason to believe that the product evaluation should be considered negative. In context, this is what the reviewer actually claimed about the product:

(9) 40 euros que cuesta... Vamos, toda una ganga teniendo en cuenta la magnifica calidad del producto!! Hombres del planeta Tierra, este es el regalo perfecto para vuestras novias/esposas/amantes... Regalad Fantasy y recibireis un buen... guantazo!!

The use of words such as "ganga", "magnífica calidad" and "regalo perfecto" turn out to be *ironic* in context. However texts that use irony account for a minority of cases, thus not affecting our analysis results considerably.

One last aspect regarding the content of the reviews that seems to affect the degree of success in Sentitext's ratings was already mentioned in section 2.1, namely, the difference between reviews pertaining to products that have some conceptual content themselves (movies, books, etc.) and those that do not. In some cases, the negative words included in the texts relate not to the reviewer's opinion, but rather to the conceptual content of what is being reviewed. With reference to the texts in which the average human rating is two points higher than that of Sentitext and triggered a change in polarity (only 3%), those texts discuss films whose themes are horror and witchcraft, and include a detailed account of the plot, which explains the large number of negative words.

7 Conclusions

The results obtained can clearly be considered as outstanding in general terms. Better accuracy is achieved for reviews of products/artifacts rather than movies or books, where an account of the conceptual content may include affect-loaded words and therefore interferes with the overall result.

Another source of trouble for our knowledge-based rating inference system is irony, a high-level discourse resource very difficult to identify automatically.

We do believe, however, that accounting for other, more easily identifiable discourse markers of opinion could help improve results in a number of cases. For example, increasing the weight of negative or positive segments appearing in the opening or closing paragraphs, as done by Devitt and Ahmad (2007), seems reasonably feasible and could produce better results.

References

Akkaya, C. & Wiebe, J., 2009. Subjectivity word sense disambiguation. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1*. Singapore: Association for Computational Linguistics, pp. 190-199.

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

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

Balahur, A., Kozareva., & Montoyo, A., 2009. Determining the Polarity and Source of Opinions Expressed in Political Debates. in Proceedings of Tenth International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2009), Mexico City, Mexico, March .

Boldrini, E., Fernández, J.,Gómez, J.M., Martínez-Barco, P., 2009. Machine Learning

Techniques for Automatic Opinion Detection in Non-Traditional Textual Genres. In: 1st Workshop on Opinion Mining and Sentiment Analysis (WOMSA), 13 Nov 2009, Sevilla, Spain.

Cruz, F.L., Troyano, J.A., Enríquez, F., Ortega, F.J., 2008. Clasificación de documentos basada en la opinión: Experimentos con un corpus de de críticas de cine en español. Boletín de la Sociedad Española para el Procesamiento del Lenguaje Natural.

Devitt, Ann & Ahmad, K., 2007. Sentiment analysis in financial news: A cohesion-based approach. In *Proceedings of the Association for Computational Linguistics (ACL).* pp. 984-991.

Esuli, A. & Sebastiani, F., 2006. SENTIWORDNET: A Publicly Available Lexical Resource for Opinion Mining. In *Proceedings of the 5th Conference on Language Resources and Evaluation (LREC '06).* LREC 2006. pp. 417-422.

Godbole, N., Srinivassaiah, M. & Skiena. S., 2007. Large-scale sentiment *analysis* for new and blogs. In Proceedings of the International Conference on Weblogs and Social Media (ICWSM).

Goldberg, A.B. & Zhu, X., 2006. Seeing stars when there aren't *many* stars: graph-based semi-supervised learning for sentiment categorization. In *Proceedings of the First Workshop on Graph Based Methods for Natural Language Processing*. TextGraphs-1. Stroudsburg, PA, USA: Association for Computational Linguistics, p. 45–52.

Gupta, N., Di Fabbrizio, G. & Haffner, P., 2010. Capturing the stars: predicting ratings for service and product reviews. In *Proceedings of the NAACL HLT 2010 Workshop on Semantic Search*. SS '10. Stroudsburg, PA, USA: Association for Computational Linguistics, p. 36–43.

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

Kim, S.-M. & Hovy, E., 2004. Determining the sentiment of opinions. In *Proceedings of the 20th international conference on Computational Linguistics*. Geneva, Switzerland: Association for Computational Linguistics, p. 1367.

Moreno-Ortiz, A., Pérez Pozo, Á. & Torres Sánchez, S., 2010a. Sentitext: sistema de análisis de sentimiento para el español. *Procesamiento de Lenguaje Natural*, 45, p.297-298.

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

Pang, B. & Lee, L., 2004. A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*. Barcelona, Spain: Association for Computational Linguistics, p. 271.

Pang, B. & Lee, L., 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of ACL 2005*. ACL. pp. 115-124.

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

Riloff, E., Patwardhan, S. & Wiebe, J., 2006. Feature subsumption for opinion analysis. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*. EMNLP '06. Stroudsburg, PA, USA: Association for Computational Linguistics, p. 440–448.

Riloff, E., Wiebe, J. & Wilson, T., 2003. Learning subjective nouns using extraction pattern bootstrapping. In *Proceedings of the seventh conference on* Natural *language learning at HLT-NAACL 2003 - Volume 4*. CONLL '03. Stroudsburg, PA, USA: Association for Computational Linguistics, p. 25–32.

Shimada, K. & Endo, T., 2008. Seeing several stars: a rating inference *task* for a document containing several evaluation criteria. In

Proceedings of the 12th Pacific-Asia conference on Advances in knowledge discovery and data mining. PAKDD'08. Berlin, Heidelberg: Springer-Verlag, p. 1006–1014.

Taboada, M. & Grieve, J., 2004. Analyzing Appraisal Automatically. In *AAAI Technical Report SS-04-07*. American Association for Artificial Intelligence Spring Symposium on Exploring Attitude and Affect in Text. Stanford, pp. 158-161. Available at: http://www.sfu.ca/~mtaboada/docs/TaboadaGrieveAppraisal.pdf.

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

Wiebe, J. & Mihalcea, R., 2006. Word sense and subjectivity. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual* meeting *of the Association for Computational Linguistics*. Sydney, Australia: Association for Computational Linguistics, pp. 1065-1072.