Three-class Sentiment Analysis adapted to short texts

Análisis del sentimiento en tres clases adaptado a textos cortos

Alexandre Trilla, Francesc Alías
GTM – Grup de Recerca en Tecnologies Mèdia
LA SALLE – UNIVERSITAT RAMON LLULL
Quatre Camins 2, 08022 Barcelona (Spain)
atrilla@salle.url.edu, falias@salle.url.edu

Resumen: La demanda de información sobre opiniones y sentimiento se ha incrementado los últimos años. Este artículo adapta un sistema general de análisis del sentimiento para textos cortos y tres clases de sentimiento. Se identifica el sentimiento positivo, negativo y neutro de forma automática con técnicas de Ingeniería de atributos y Clasificación de Texto. Para evaluar la efectividad de este esquema se utiliza el conjunto de datos Semeval 2007, con el que se alcanza una tasa máxima del 49%, mejorando un 7% los resultados presentados en el estado del arte siguiendo las mismas condiciones de evaluación.
Palabras clave: Análisis del Sentimiento, Aprendizaje Computacional, Ingeniería de atributos, Clasificación de Texto

Abstract: The demand for information on opinions and sentiment has seen an increase in recent years. This article adapts a general Sentiment Analysis scheme to deal with short texts and three classes of sentiment. It addresses positive, negative and neutral sentiments automatically using Feature Engineering and Text Classification techniques. The effectiveness of this scheme is evaluated using the Semeval 2007 dataset and it achieves maximum rate of 49%, improving by 7% the results reported in the state of the art following the same evaluation conditions.
Keywords: Sentiment Analysis, Machine Learning, Feature Engineering, Text Classification

1 Introduction

In the recent years the field of Sentiment Analysis (SA) has experienced a substantial raise in response to the surge of interest in affective computing for inferring knowledge and understanding from people's opinions, e.g., in social networks, marketing 2.0, etc. The detection of sentiment in text can be conceived as an expert heuristic process where the specific knowledge is hard-coded into the system via a set of rules, or else as an automatic induction process based on Machine Learning (ML) that discovers it from available user data (Strapparava y Mihalcea, 2007; Pang y Lee, 2008). With regard to the latter data-driven approach, the goal is to maximise the classification effectiveness through delving into the linguistic parameters and the language models that can be extracted from the text of analysis (Pang y Lee, 2008; Strapparava y Mihalcea, 2007; Dang, Zhang, y Chen, 2010). The features of use often exploit the n-gram representation of text in a weighted vector space (essentially unigrams and bigrams), and sometimes they are represented along with their Part-Of-Speech (POS) tags, stems, etc. (Pang y Lee, 2008; Dang, Zhang, y Chen, 2010). Then, the features extracted from the text are operated with diverse classifiers such as Multinomial Naive Bayes, Maximum Entropy and Support Vector Machine (Pang, Lee, y Vaithyanathan, 2002; Pang y Lee, 2008). Moreover, these conventional SA solutions are usually set to work with big compilations of long texts labelled with two opposite sentiment categories, e.g., full product reviews of positive and negative opinions that may amount up to about 55000 sentences, for example (Pang, Lee, y Vaithyanathan, 2002). However, there are other potential applications that operate in different settings (e.g., with short texts and/or three classes of sentiment), thus requiring an adapted version of this general design. For instance, see the works on social media mining with Twitter

(Kouloumpis, Wilson, y Moore, 2011), fairy tales (Alm, Roth, y Sproat, 2005) and Text-To-Speech synthesis (Alías et al., 2008).

Following (Strapparava y Mihalcea, 2007), the present work focuses on the latter kind of applications, specifically, the positive, negative and neutral sentiment categorisation of short texts. In this setting, the granularity of the text under analysis is usually determined to be the sentence, as sentences are sensibly short textual representations with a rich affective content, allowing natural expressive variations between them within the same paragraph (Alm, Roth, y Sproat, 2005; Alías et al., 2008). To train and evaluate the effectiveness of the SA schemes adapted to this scenario, the Semeval 2007 dataset is of use (Strapparava y Mihalcea, 2007). The motivation for considering these data is two-fold: 1) the corpus provides the three-class sentiment labelling produced and validated by human evaluators, and 2) the affective features (if present) are guaranteed to appear in these short sentences (i.e., news headlines) (Strapparava y Mihalcea, 2007), in contrast to long texts where a single label corresponds to the overall sentiment wash (Strapparava y Mihalcea, 2007). This work performs a SA process along the lines of (Pang, Lee, y Vaithyanathan, 2002; Pang y Lee, 2008) but in a different scenario: short texts and three sentiment classes. Its main purpose is to address the SA problem at the sentence level with techniques that are usually effective at the document level. It focuses on determining the relevant features of use and evaluates the adaptation of several classifiers to the problem at hand. Finally, it compares the obtained results with the state of the art (Strapparava y Mihalcea, 2007) to determine the strategy that most effectively fits the problem.

The paper is organised as follows. Section 2 presents the learning details of the ML classification approaches that are typically used in SA. Section 3 describes the experiments and analyses the obtained results. Section 4 discusses the resulting effectiveness rates and draws the conclusions of this work.

2 Adaptation of Sentiment Analysis to three classes and short texts

This section focuses on the relevant features of use, whether they be unigrams alone or a full set of linguistic and affective features (Pang, Lee, y Vaithyanathan, 2002; Dang, Zhang, y Chen, 2010), i.e., the Feature Engineering, and the adequate Text Classification (TC) strategies, which may infer a generative model of fit a discriminating function (Pang, Lee, y Vaithyanathan, 2002), see Figure 1.

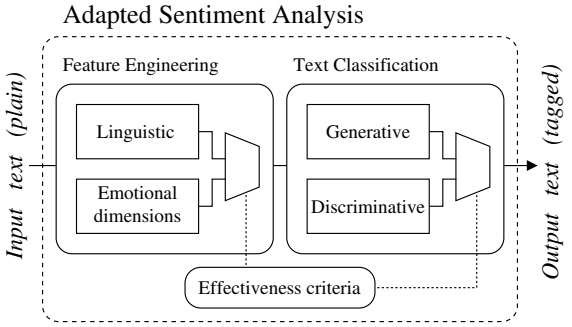

Figure 1: Block diagram of the proposed Sentiment Analysis approach, considering both the diversity in the nature of the features extracted from the text and the diversity in the learning principles of the classifiers.

2.1 Feature Engineering

A front-end task to the actual classification of sentiment is the modelling and uniform representation of the features. To that end, the Vector Space Model (VSM) representation is used, which shapes the input text as a vector with one real-valued component for each feature (Sebastiani, 2002; Manning, Raghavan, y Schütze, 2008).

In general, it is the semantics which provide a great deal of information with respect to the affect in text (Pang y Lee, 2008). This essentially leads to modelling *words*, which are plausibly conceived to be the smallest meaningful units of affect (Batliner et al., 2009). Words alone, which are modelled as unigrams, are obtained from the lexical instances of the tokens. Their consideration in isolation constitutes a simple Bag-Of-Words (BOW) model, which does not account for the order of words appearing in a text (Sebastiani, 2002). In certain contexts, this BOW model is regarded to be the most adequate model (Pang, Lee, y Vaithyanathan, 2002). In other contexts, it simply lacks useful information (Alías et al., 2008). For the latter cases, it is often useful to increase the number of features by considering patterns that are particularly discriminative (Manning, Raghavan, y Schütze, 2008).

In this regard, bigrams (i.e., the ordered co-occurrence of two unigrams) may also be considered in the amount of features (Pang, Lee, y Vaithyanathan, 2002). Bigrams are reported to be of help to grasp stylistic traits and structural information (i.e., syntactic) in the text (Alías et al., 2008; Pang y Lee, 2008). This is regarded to be an alternative way to incorporate context (Pang, Lee, y Vaithyanathan, 2002), and with the inclusion of POS tags, the analysis is added some grammatical and syntactical value (Pang y Lee, 2008). Nevertheless, higher order n-grams are generally discarded as they do not appear to contribute much to the identification of affect in the text (Pang y Lee, 2008). In addition, the stems of the words may also be considered for enhanced indexing purposes (Sebastiani, 2002), and a semantic expansion procedure may also be conducted through the inclusion of word synonyms (García y Alías, 2008). Finally, non-linguistic traits may also be considered as a means of domain independent features. In this regard, the emotional dimensions of valence, activation and control are usually considered (García y Alías, 2008; Trilla y Alías, 2009).

2.1.1 Term Weighting

In TC, the relative importance of features is of great relevance (Sebastiani, 2002; Manning, Raghavan, y Schütze, 2008). But using all the features together directly often increases the the size of the feature space without providing much satisfactory power (sparseness problem) (Manning, Raghavan, y Schütze, 2008). Hence, weighting the relevance of the features increases the separability properties of the data improving the classification effectiveness (Sebastiani, 2002; Manning, Raghavan, y Schütze, 2008; Dang, Zhang, y Chen, 2010).

An everlasting question regarding the weighting of terms is their representation of presence versus frequency (Pang, Lee, y Vaithyanathan, 2002; Pang y Lee, 2008; Manning, Raghavan, y Schütze, 2008). Although the frequency of terms seems to be more useful as it naturally encodes the presence of terms, the use of binary weights denoting term presence/absence has comparatively performed better in SA (Pang y Lee, 2008). In this work, binary weights are evaluated, as well as a couple of enhanced frequency-based weights: the Inverse Term Frequency (ITF) (Alías et al., 2008), which weights each term according to its prominence within the sentence, and the Relevance Factor (RF) (Lan et al., 2009), which weights the relevance of a term regarding its distribution among the categories.

2.2 Text Classification

This section describes some of the most representative TC methods for SA, focusing on the discovery of knowledge that each method can provide from the input features. Given the short text conditions tackled in this work, the choice of classifier probably has an important effect on the effectiveness of the system (Manning, Raghavan, y Schütze, 2008).

Originally the classification step was performed with a set of heuristic rules on the circumplex (García y Alías, 2008), but recent improvements have shown that automatically learning the term-feature space is a more effective solution (Trilla, Alías, y Lozano, 2010). Hence, to capture the generality and scope of the problem space, both generative and discriminative learning approaches are considered in this work (see Figure 1). Generative models explain the data, and if the model is correct, they should yield the best possible classification effectiveness rates (Mitchell, 2005). Nevertheless, since the form of the actual model is unknown and the training sample does not generally cover the whole feature space, instead of proposing an endless amount of possible approximate models, task-centric approaches based on discriminating the sentiment categories are evaluated (Manning, Raghavan, y Schütze, 2008).

In the end, the inductive construction of ML methods for solving TC and SA is essentially the same. Within the polynomial models, linear models are proposed in this work for their simplicity over their (more complex) nonlinear counterparts. Note that because of the bias-variance tradeoff in the classification effectiveness rates, complex models are not systematically better than linear models (Manning, Raghavan, y Schütze, 2008). Nonlinear models have more parameters to fit on a limited amount of training data and they are more prone to make mistakes for small datasets (see (Alías et al., 2008) for an empirical evidence of this phenomenon). Instead, linear models might be preferable to separate the bulk of the data, i.e., to obtain a better generalisation of classification (Manning, Raghavan, y Schütze, 2008). And

with the high dimensional spaces that are typically encountered in text processing applications, the likelihood of linear separability increases rapidly (Manning, Raghavan, y Schütze, 2008). What follows is the description of some typical learning environments in TC and SA to evaluate.

2.2.1 Multinomial Naive Bayes (MNB)

MNB is a probabilistic generative approach that builds a language model assuming conditional independence among the features. In reality, this assumption does not hold for text data (Pang, Lee, y Vaithyanathan, 2002), but even though the probability estimates are of low quality because of this oversimplified model, its classification decisions are surprisingly good (Manning, Raghavan, y Schütze, 2008). The MNB combines efficiency (it has an optimal time performance) with good accuracy, hence it is often used as a baseline in TC and SA research (Sebastiani, 2002; Manning, Raghavan, y Schütze, 2008).

2.2.2 Associative Relational Network - Reduced (ARN-R)

It is a word co-occurrence network-based approach that constructs a VSM with a term selection method "on the fly" based on the observation of test features (Alías et al., 2008). This inherent term selection refinement is reported to improve the classical VSM for modest-size sentence-based data (Alías et al., 2008). Dense vectors representing the input text and the class are retrieved (no learning process is involved) and evaluated by the cosine similarity measure. The basic hypothesis in using the ARN-R for classification is the contiguity hypothesis, where terms in the same class form a contiguous region, and regions of different classes do not overlap (Manning, Raghavan, y Schütze, 2008).

2.2.3 Latent Semantic Analysis (LSA)

LSA is similar to the VSM, but builds a latent semantic space by computing the Singular Value Decomposition (SVD) of the term-class matrix obtained from the VSM (i.e., constructing a low-rank approximation with its principal eigenvectors) (Manning, Raghavan, y Schütze, 2008). The cosine similarity between the class vectors and the query text vectors (obtained by adding the observed term vectors) is used to make decisions in the reduced latent space. LSA has been used for affect classification (Bellegarda, 2011) as well as in TC and SA (Sebastiani, 2002; Strapparava y Mihalcea, 2007).

2.2.4 Maximum Entropy (MaxEnt)

It is a probabilistic discriminative approach that fits a set of exponential functions via the Maximum A Posteriori estimation (Carpenter, 2008). MaxEnt obeys the maximum entropy principle, therefore it does not make any further assumption beyond what is directly observed in the training data. Moreover, it makes no assumptions about the relationships among the features, and so might potentially be more effective when conditional independence assumptions are not met (Pang, Lee, y Vaithyanathan, 2002). MaxEnt has been used for SA and TC environments (Trilla, Alías, y Lozano, 2010; Pang, Lee, y Vaithyanathan, 2002; Pang y Lee, 2008)

2.2.5 Support Vector Machine (SVM)

It is a maximum-margin discriminative approach that searches the hyperplane (decision surface in the feature space) that is maximally distant from the class-wise data points. Since the SVM is a dichotomous classifier, a multicategorisation strategy has to be considered to deal with the three sentiment classes. SVM has shown to be superior with respect to other methods in situations with few training data (Pang y Lee, 2008), in TC scenarios (Sebastiani, 2002; Lan et al., 2009) as well as in SA (Pang, Lee, y Vaithyanathan, 2002; Pang y Lee, 2008).

3 Empirical evaluation

To evaluate and determine the SA strategy that yields the best effectiveness in identifying the sentiment in short texts for the problem at hand, the dataset of use in this work is the Semeval 2007 (Strapparava y Mihalcea, 2007), for its convenience to address the three-category sentiment analysis at sentence level (Strapparava y Mihalcea, 2007). It consists of a compilation of news headlines (taken for short sentences with less than 8 words on average) drawn from major newspapers. Its design criteria highlight its typically high load of affective content written in a style meant to attract the attention of the readers (Strapparava y Mihalcea, 2007). In addition, its short-text form is adequate as a single label represents the whole sentence (Alías et al., 2008), whereas in long texts, the

Instance properties	Counts	
Total (sentences)	1250	
Positive	174	
Neutral	764	
Negative	312	
With repeated words	46	
Idem without stop words	4	
Average length	7.53	
Feature properties	**Unig.**	**Big.**
Total (n-grams)	8115	6865
Vocabulary	4085	6251
Frequent ($\geq$5)	226	14

Table 1: Properties of the Semeval 2007 dataset in terms of instance and feature counts.

labelling may hide a sentiment wash (Pang, Lee, y Vaithyanathan, 2002). This corpus is distributed in two sets: one for trial (containing 250 headlines) and the other for testing (containing 1000 headlines).

3.1 Preliminary analysis of the corpus

An overall description of the properties of the entire dataset is shown in Table 1. Note that the amount of sentences (i.e., instances) in the corpus with words appearing more than once in a single sentence is small (46 sentences out of 1250 yield a rate of 3.68%), and this figure even drops more if stop words are filtered out (0.32%). This fact shows that differentiating between the presence/frequency representation of the features seems to be of little relevance for this data: in either case, the information is almost the same (this is strictly true for the 99.68% of the sentences in this corpus).

It is also important to note the richness of the vocabulary extracted from the data. Half the total number of unigrams yields the size of the whole unigram set, and in the case of bigrams, these counts are almost equal. Hence, on average, each term only appears twice at most in the whole corpus. This lack of frequent features puts an extra difficulty for the identification of sentiment and therefore supports the proposal of Feature Engineering on the most relevant ones.

In order to gain intuition of the data character, Table 2 shows the relative balance of some word counts among the sentiment classes. As the orientation of the words changes from "good" to "bad", the mass of the word counts shifts from the positive sentiment to the negative sentiment. This fact reveals the relevance of certain words as sentiment indicators and shows what the classifiers may eventually learn from the data.

Orient.	Word	Pos.	Neu.	Neg.
Good	sweet	2	0	0
	record	10	1	0
Fine	good	5	6	0
	help	4	7	1
Fair	talk	3	10	1
	say	3	23	9
Mean	fail	0	2	2
	crash	0	3	7
Bad	fear	0	1	4
	dead	0	3	12

Table 2: Balance of word counts among the sentiment classes with respect to an orientation grading from "good" to "bad". The strength of cell shading denotes the mass of word counts.

3.2 Experimental results

The SA approaches under evaluation are described as follows. On the one hand, the features of use contrast two approaches (Pang y Lee, 2008): 1) the sensible agglomeration of traits that are reported to be useful for SA, i.e., weighted unigrams, bigrams, POS tags, stems, emotional dimensions and negation flags, and 2) the sole consideration of weighted unigrams as only the essential traits of sentiment in text. On the other hand, the specific implementation of the TC methods to be evaluated are described hereunder:

- MNB uses Manning's TC definition for discrete features (binary weights) (Manning, Raghavan, y Schütze, 2008) and the Weka's general-purpose NaiveBayesMultinomial with continuous weighted features (Witten y Frank, 2005).
- ARN-R is implemented following (Alías et al., 2008).
- LSA uses the SVD implementation provided by LingPipe[1] to construct a latent semantic space (Deerwester et al., 1990).
- MaxEnt uses the Stochastic Gradient Descent optimisation procedure provided by LingPipe (Carpenter, 2008).

[1] http://alias-i.com/lingpipe/

- SVM uses the Weka's Sequential Minimum Optimisation with a linear kernel and pairwise classification (Witten y Frank, 2005).

In (Strapparava y Mihalcea, 2007), one of the effectiveness rates used for the evaluation of the classification strategies was the macroaveraged F_1 measure, which is also customary to use in TC (Sebastiani, 2002; Manning, Raghavan, y Schütze, 2008). This unweighted effectiveness measure is needed to even the importance of each class regardless of the corpus instance imbalances, see Table 1. Note that the size of the neutral class is more than four times bigger that the size of the positive class, i.e., the class with the least generality, which makes it more difficult to effectively model the latter smaller class.

As far as we know, the best F_1 result published in the state of the art for sentiment classification with the Semeval 2007 corpus is set at 42.43% (Strapparava y Mihalcea, 2007). This effectiveness rate was obtained with a Naive Bayes classifier, predicting a valence score for the sentiment, and overtrained with additional data that was manually annotated with positive and negative sentiments. This section studies if the methodology proposed in this work provides a more effective system for the problem at hand. The comparison with respect to the state of the art entails evaluating the effectiveness of the system with a train-test scenario (Strapparava y Mihalcea, 2007), where a single F_1 measure is provided given that only one experiment is performed (training with the trial subset of the corpus that consists of 250 headlines, and testing with the remaining 1000 headlines).

The effectiveness of the classifiers with the whole set of features (weighted unigrams, bigrams, POS tags, stems, emotional dimensions and negation flags) is shown in Table 3. It can be observed that most of them yield similar effectiveness rates around 39%, so none of them improves the aforementioned baseline result in the literature. In addition, MaxEnt could not predict the class with the least generality, which denotes the especial requirement of a minimum amount of examples for this classifier. Regarding that the feature dimensionality is very large in this scenario (considering all unigrams and bigrams together amount up to more than 10000 parameters), it is possible that the classifiers overfit the training data, therefore not performing properly. Overfitting generally occurs when a model is excessively complex, such as having too many parameters relative to the number of training instances (Sebastiani, 2002; Manning, Raghavan, y Schütze, 2008).

Whole set of features

Classifier	**Term Weighting**		
	Binary	ITF	RF
MNB	40.26	42.20	N/A
ARN-R	37.38	33.40	39.36
LSA	33.44	34.81	30.26
MaxEnt	N/A	N/A	N/A
SVM	39.27	37.76	38.94

Table 3: F_1 results with the whole set of features: weighted unigrams, bigrams, POS tags, stems, emotional dimensions and negation flags. N/A stands for Not Available due to not predicting the class with the least generality.

In this regard, this work also experiments with weighted unigrams alone, thus grasping only the essence of the sentiment in this short text. The results with unigrams alone are shown in Table 4. The reduced feature setting enables the classifiers to generalise better (Manning, Raghavan, y Schütze, 2008), and this reveals three classifiers that improve the baseline effectiveness rate at least by 2%: the MNB, MaxEnt and SVM. Specifically, the MNB with binary-weighted unigrams and the MaxEnt with RF yield the best improvement margin, which is of 7%. In this lighter but essential feature setting, which involves much less parameters, the classifiers perform more effectively, a fact that is attributed to minimising the overfitting of the data (Sebastiani, 2002; Manning, Raghavan, y Schütze, 2008). Hence, they yield a good adaptation of the general SA methods to the problem at hand. In (Pang, Lee, y Vaithyanathan, 2002), a similar outcome was obtained with respect to the importance of unigrams alone for long texts labelled with two categories of sentiment.

In the end, the most successful SA strategies evaluated for the problem at hand, namely with MNB and MaxEnt, converge to a similar effectiveness around 49%, thus improving the effectiveness rates reported in the state of the art by almost 7% (Strapparava y Mihalcea, 2007).

Unigram features

Classifier	Term Weighting		
	Binary	ITF	RF
MNB	48.89	45.41	N/A
ARN-R	37.26	32.32	42.25
LSA	37.71	37.63	31.96
MaxEnt	N/A	N/A	49.26
SVM	45.30	36.83	N/A

Table 4: F_1 results with plain unigram features. N/A stands for Not Available due to not predicting the class with the least generality.

4 Discussion and Conclusions

The identification of affect in text is a complex problem that has many facets to consider. In this work, we have intended to perform an exhaustive and comprehensive study to tackle a particular three-class sentiment analysis problem at the sentence level framed by a small dataset, which is the Semeval 2007 dataset (Strapparava y Mihalcea, 2007). Our experiments indicate that under such problem settings, the success of a good classifier such as MNB or MaxEnt depends on the representation of the features, which helps the classifier to not overfit the data (Manning, Raghavan, y Schütze, 2008). In fact, overfitting may be reduced if the number of training examples is roughly proportional to the number of features used to represent the data (Sebastiani, 2002). This work shows how considering unigrams alone (with adequate weighting methods) results in better classification effectiveness compared to using additional features such as bigrams, POS tags, etc. Previous works operating in other environments, namely longer texts and two classes of sentiment, reached a similar conclusion with regard to the importance of unigrams (Pang, Lee, y Vaithyanathan, 2002). These results allow us to suggest that for SA problems, using only the essential information that denotes the sentiment in text by means of the unigrams alone, the problem becomes more tractable for the generally successful classifiers, and therefore they performs most effectively.

Bibliografía

Alías, Francesc, Xavier Sevillano, Joan Claudi Socoró, y Xavier Gonzalvo. 2008. Towards High-Quality Next-Generation Text-to-Speech Synthesis: A Multidomain Approach by Automatic Domain Classification. *IEEE Trans. Audio, Speech, Lang. Process.*, 16(7):1340–1354, Sep.

Alm, Cecilia Ovesdotter, Dan Roth, y Richard Sproat. 2005. Emotions from text: machine learning for text-based emotion prediction. páginas 579–586.

Batliner, Anton, Dino Seppi, Stefan Steidl, y Björn Schuller. 2009. Segmenting into adequate units for automatic recognition of emotion-related episodes: a speech-based approach. *Advances in Human Computer Interaction (AHCI)*.

Bellegarda, J.R. 2011. A Data-Driven Affective Analysis Framework Toward Naturally Expressive Speech Synthesis. *IEEE Trans. Audio, Speech, Lang. Process.*, 19(5):1113–1122, Jul.

Carpenter, Bob. 2008. Lazy Sparse Stochastic Gradient Descent for Regularized Multinomial Logistic Regression. Informe técnico, Alias-i, Inc.

Dang, Yan, Yulei Zhang, y Hsinchun Chen. 2010. A Lexicon-Enhanced Method for Sentiment Classification: An Experiment on Online Product Reviews. *IEEE Intell. Syst.*, 25(4):46–53, Jul.-Aug.

Deerwester, Scott, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, y Richard Harshman. 1990. Indexing by Latent Semantic Analysis. *J. Am. Soc. Inform, Sci.*, 41(6):391–407.

García, David y Francesc Alías. 2008. Emotion identification from text using semantic disambiguation. En *Procesamiento del Lenguaje Natural*, numero 40, páginas 75–82 *(in Spanish)*, Mar.

Kouloumpis, Efthymios, Theresa Wilson, y Johanna Moore. 2011. Twitter Sentiment Analysis: The Good the Bad and the OMG! Jul.

Lan, Man, Chew Lim Tan, Jian Su, y Yue Lu. 2009. Supervised and Traditional Term Weighting Methods for Automatic Text Categorization. *IEEE T. Pattern. Anal.*, 31(4):721–735, Apr.

Manning, Christopher D., Prabhakar Raghavan, y Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press, Cambridge, MA, USA.

Mitchell, Tom M. 2005. Generative and Discriminative Classifiers: Naive Bayes and Logistic Regression. *Online draft*, 755:1–17.

Pang, Bo y Lillian Lee. 2008. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135.

Pang, Bo, Lillian Lee, y Shivakumar Vaithyanathan. 2002. Thumbs up? sentiment classification using machine learning techniques. En *Proc. of EMNLP'02*, páginas 79–86, Philadelphia, PA, USA, Jul.

Sebastiani, Fabrizio. 2002. Machine learning in automated text categorization. *ACM Comput. Surv.*, 34:1–47.

Strapparava, Carlo y Rada Mihalcea. 2007. SemEval-2007 Task 14: Affective Text. Jun.

Trilla, Alexandre y Francesc Alías. 2009. Sentiment classification in English from sentence-level annotations of emotions regarding models of affect. páginas 516–519, Sep.

Trilla, Alexandre, Francesc Alías, y Isaac Lozano. 2010. Text classification of domain-styled text and sentiment-styled text for expressive speech synthesis. páginas 75–78, Nov.

Witten, Ian H. y Eibe Frank. 2005. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, San Francisco, CA, USA.