28 Generating a Spanish affective dictionary with supervised learning techniques

Daniel Bermudez-Gonzalez[1], Sabino Miranda-Jiménez[2], Raúl-Ulises García-Moreno[3], and Dora Calderón-Nepamuceno[4]

Abstract

Nowadays, machine learning techniques are being used in several Natural Language Processing (NLP) tasks such as Opinion Mining (OM). OM is used to analyse and determine the affective orientation of texts. Usually, OM approaches use affective dictionaries in order to conduct sentiment analysis. These lexicons are labeled manually with affective orientation (polarity) of words such as positive or negative. There are few dictionaries of affective orientation for Spanish; also, the size of these dictionaries is small. Thus, we propose a method for building a large affective Spanish dictionary for subjectivity and sentiment analysis. Supervised learning techniques are used to classify the entries from a lexical dictionary according to their affective orientations based on their definitions. We combine three classifiers (decision trees, naive Bayes, and a support vector machine) to determine the final polarity of each entry, that is, positive or negative.

Keywords: opinion mining, subjectivity and sentiment analysis, affective orientation, polarity detection.

[1] Prospectiva en Tecnología e Integradora de Sistemas SA de CV, Ciudad de México, Mexico; danielbermudez30@gmail.com

[2] INFOTEC- Centro de Investigación e Innovación en Tecnologías de la Información y Comunicación, Aguascalientes, México; sabino.miranda@infotec.mx

[3] Services & Processes Solutions, Ciudad de México, México; raulgm00@gmail.com

[4] UAEM UAP Nezahualcóyotl, Estado de México, México; dmcalderonn@uaemex.mx

How to cite this chapter: Bermudez-Gonzalez, D., Miranda-Jiménez, S., García-Moreno, R.-U., & Calderón-Nepamuceno, D. (2016). Generating a Spanish affective dictionary with supervised learning techniques. In A. Pareja-Lora, C. Calle-Martínez, & P. Rodríguez-Arancón (Eds), *New perspectives on teaching and working with languages in the digital era* (pp. 327-338). Dublin: Research-publishing.net. http://dx.doi.org/10.14705/rpnet.2016.tislid2014.445

1. Introduction

In recent years, the automatic processing of opinions has increased because of its potential applications. One of them is sentiment analysis (Pang & Lee, 2008) in social networks. Most people write their opinions in forums, review sites, and microblogging (Twitter, Facebook, among others). This information is useful for companies, governments, and individuals who want to obtain global feedback for their activities or products.

Machine learning techniques have been used to face sentiment analysis problems, namely, to determine the affectivity of texts: their positive or negative orientation. As stated by Banea, Mihalcea, and Wiebe (2011), "[m]uch of the research work […] on sentiment and subjectivity analysis has been applied to English, but work on other languages is [a] growing [need]" (p. 1).

In this paper, we propose a new method to build a subjectivity and sentiment dictionary for Spanish based on the definitions from an explanatory dictionary and three classifiers, which will be employed to perform sentence level sentiment classification.

2. Related work

Lexicons have been used for subjectivity and sentiment analysis because they can be applied to identify opinions or emotions by means of rule-based opinion classifiers. For example, there are several popular lexicons for subjectivity and sentiment analysis for English, such as the OpinionFinder lexicon (Wiebe & Riloff, 2005), which contains 6,856 unique entries associated with a polarity label (positive, negative, neutral). SentiWordNet (Esuli & Sebastiani, 2006) is another popular lexicon, which is based on WordNet (Miller, 1995) and encompasses more than 100,000 words. It was automatically generated, starting with a small set of manually labeled *synsets*. A *synset* represents a group of cognitive synonyms (nouns, verbs, adjectives, or adverbs) that express a distinct concept.

In the case of the Spanish language, there are few dictionaries of affective orientation. One of them, the Spanish Emotion Lexicon (SEL), has 2,038 words (Díaz-Rangel, Sidorov, & Suárez-Guerra, 2014; Sidorov et al., 2012). This dictionary was manually classified into 6 affective categories (joy, anger, fear, sadness, surprise, and disgust). The Polarity Lexicon (PL) presented in Saralegi and San Vicente (2013) has 4,738 words classified as positive or negative. Another lexicon, the Spanish Sentiment Lexicon (Pérez-Rosas, Banea, & Mihalcea, 2012) uses a cross-language expansion approach based on WordNet to determine the polarity. This lexicon has 3,843 words, classified as positive or negative.

Our approach is different: we used the entry definition of an explanatory dictionary to determine the polarity of the entry itself. We used two affective dictionaries manually labeled (SEL and PL) to train the classifiers in order to classify the entries from a large explanatory dictionary.

3. Building the affective dictionary

We were interested in discovering the positivity or negativity of dictionary entries in order to use them in opinion mining tasks. Thus, our objective was to automatically build an affective dictionary for Spanish with two categories: positive and negative. However, other works have used more categories to determine semantic orientation of messages or documents.

The number of classes used depends on the particular purposes and the domains of these works. For instance, Pang, Lee, and Vaithyanathan (2002) used two classes (positive and negative) for movie reviews, and Pérez-Rosas et al. (2012) used these two classes for generating sentiment lexicons in a target language using annotated English resources. Three classes (positive, negative, and neutral) were the base to predict contextual polarity of subjectivity phrases in a sentence in Agarwal, Biadsy, and Mckeown (2009); four classes (positive, negative, neutral, and informative) to determine the semantic orientation on Twitter data (Sidorov et al., 2012); and six classes helped determine a fine-grained affective orientation in sentences (joy, anger, fear, sadness, surprise, and disgust) in Díaz-Rangel et al. (2014).

In our approach, we used three dictionaries in order to obtain the resulting affective dictionary: an explanatory dictionary, which has the words (entries) to be classified, and two lexicons, labeled by hand with different affective categories.

3.1. The affective lexicons

Two labeled affective dictionaries were used to train the classifiers: the SEL lexicon (Díaz-Rangel et al., 2014; Sidorov et al., 2012) and the PL lexicon (Saralegi & San Vicente, 2013). In our approach, we used only two categories; thus, we mapped SEL's categories to a positive or negative category, that is, *joy* and *surprise* to positive, *anger*, *fear*, *sadness*, and *disgust* to negative.

3.2. Preprocessing of the explanatory dictionary

We used the words from Anaya explanatory dictionary of Spanish as input data to be classified (30,228 entries). Also, we used the entry definitions in two different ways: first, the definitions of words from the affective dictionaries were used to train the classifiers, and second, the definitions of the remaining entries were used to classify the entry itself.

In order to prepare the dictionary entries for classification, we removed from the entries all phrases and words with no alphabetic symbols, suffixes, and prefixes (such as *mountain bike, modus vivendi*, '*ido, ida*', -*a*, or *neumo*-). Also, we removed stop words such as articles, prepositions, and conjunctions. We just used content words, that is, single words such as *abeja*, *abrumar*, *rata*, etc.

To process the definition of each dictionary entry, we applied some rules. For example, if the entry definition had a text such as '**véase** CONCEPT' (**see** CONCEPT), then the definition of the CONCEPT was searched in the explanatory dictionary and was used instead of '**véase** CONCEPT'. In the following example, the definition of *alcohómetro* is replaced by the definition of *alcoholímetro*. Applying this rule the substitution is as follows:

ENTRY	**DEFINITION**
alcohómetro	***véase*** *alcoholímetro*
alcohómetro	*Dispositivo para medir la cantidad de alcohol presente en el aire expirado por una persona.*

We also removed from definitions numbers, suffixes, prefixes, phrases with abbreviations, and abbreviations such as '*del lat.*', '*del ár.*', *FAM., vulg.*, among others. For example, this sort of particle is removed in the following definitions:

ENTRY	**DEFINITION**
ruborizar	*sonrojar, adquirir o producir rubor en el rostro.* ~~***FAM***~~*. ruborizado.*
palmar	~~***del lat.***~~ *palmare, golpear; o del caló palmar, acabar.*

Finally, in this step, if the words had multiple definitions, we used the most frequent ones, that is, we selected a percentage of the definitions, because different definitions of a same word have different affective orientations for the same word. In the following example, the word *perro* (dog) has multiple definitions and each definition has a different affective orientation; the definitions 2 and 3 have clearly negative meanings, and the definitions 5, 7, and 8 have positive meanings.

ENTRY	**DEFINITION**
perro	*1) Nombre común de cierto mamífero carnívoro, doméstico, del que hay infinidad de razas muy distintas entre sí por la forma, el tamaño y el pelaje* *2)* ***muy malo*** *[lleva una vida de perro]* *3) Dícese de la* ***persona vil, traidora y astuta*** *[no te fíes de él, es muy perro]* *4) Persona que siempre va pegada a otra.* *5)* ***Persona que acompaña tenazmente a otra para protegerla*** *de supuestos peligros.* *6) Perro faldero. Perro pequeño que siempre acompaña a su amo.* *7) Perro guardián.* ***Perro que guarda una propiedad.***

> *8) Atar los perros con longaniza. Frase con que se da a entender la* ***abundancia o riqueza.***

Because the linguistic resource that we have created aims at supporting other practical applications for opining mining, we classified the most frequent meanings of the explanatory dictionary in order to avoid too many semantic orientations of a same word. That is, we selected the first definitions up to a predetermined percentage. The explanatory dictionary lists its entries from the most frequent meanings to the least frequent meanings.

The percentage was defined using the following rules. If a word had 1-3 senses (definitions), all senses were used; if a word had 4-6 senses, 80% of the senses were used; if a word had 7-10 senses, 60% are used; if a word had more than ten senses, 40% are used. For instance, five senses were selected for the word *perro* mentioned above.

We used these percentages because the distribution of number of senses per word is substantially reduced after three senses in the Anaya dictionary. For example, words that had from one to three senses were 26,064; four senses, 1,792; and five senses, 774 (Gelbukh, Sidorov, & Ledo-Mezquita, 2003).

3.3. Preprocessing of the training data

In order to train the classifiers, we used the word definitions of SEL and PL. We used only the content words of the definitions, as we mentioned. In order to reduce their dimensionality, the Porter (2006) stemming algorithm for Spanish was applied to the definitions.

For example, after applying the preprocessing to the original text (1), we obtained a transformed text (2) which is used as a unigram model to train the classifiers, that is, we use single words (stems).

alegre dícese de la persona, gesto, etc., que tiene o denota alegría	**original text (1)**
alegr dices person gest tien denot alegr	**transformed text (2)**

3.4. Selected classifiers

Our method uses three machine learning classifiers. The selected machine learning classifiers were Naïve Bayes (NB), Decision Trees (DTs) and Support Vector Machines (SVMs).

We used the Waikato Environment for Knowledge Analysis (WEKA) software (Hall et al., 2009) that implements the machine learning algorithms mentioned above, and we implemented our version of the NB algorithm. WEKA implements SVM as a Sequential Minimal Optimisation (SMO), and DT with J48 algorithms.

Our input data is a vector. Each entry (stem) in the vector corresponds to a feature. For SVM and DT, a Term Frequency-Inverse Document Frequency (TF-IDF) weighting approach was used (Salton & Buckley, 1988), that is, we used not only the presence of each feature, but also its global importance in the explanatory dictionary.

We used a training set that consists of 5,222 words, which came from the two affective lexicons (SEL and PL). The training set has 1,924 positive words and 3,298 negative words.

Additionally, the test set consists of 3,000 words that were randomly selected from the explanatory dictionary in order to be labeled manually to assess our method. The resulting test set has 2,316 positive words and 684 negative words.

4. Experiments and results

As mentioned above, our method uses three machine learning classifiers and consists of three steps.

First, we generated the model for each classifier considering the training data. Second, we classified the entries from the explanatory dictionary using each model; thus, three affective dictionaries were generated, that is, one for each classifier. Third, we combined the results of the three classifiers in one affective dictionary using a voting scheme. For example, the word *hurgar* (delve) was labeled as positive (p) by NB, negative (n) by SVM, and negative (n) by DT; thus, the global orientation was negative because there were two votes for negative orientation. This strategy was applied for all entries from the Anaya dictionary.

We applied standard measures used in many NLP tasks. The *precision* (P) of a system is computed as the percentage of correct answers given by the automatic system. *Recall* (R) is defined as the number of correct answers given by the automatic system over the total number of answers to be given. *F-measure* is the harmonic mean of *precision* and *recall*. The coverage is 100%, thus precision and recall are equal (Navigli, 2009).

$$P= \frac{\textit{\# correct answers provided}}{\textit{\# answer provided}} \qquad R= \frac{\textit{\# correct answers provided}}{\textit{\# total answers to provide}}$$

The results obtained for each classifier are shown in Table 1. As shown in this table, we generally obtained better results when we combined the results of the classifiers. The precision obtained was 67%. The precision for the positive category was 71%, and for the negative category was 53.3%.

Table 1. Evaluation of the classifiers

Classifier	Category	Precision	Recall	F-measure
SVM	-	49.4%	49.4%	49.4%
	Positive	42.0%	42.0%	42.0%
	Negative	74.2%	74.2%	74.2%
DT	-	30.3%	30.3%	30.3%
	Positive	12.7%	12.7%	12.7%
	Negative	89.7%	89.7%	89.7%
NB	-	59.6%	59.6%	59.6%
	Positive	60.5%	60.5%	60.5%
	Negative	56.7%	56.7%	56.7%
Voting Scheme	-	**67.0%**	**67.0%**	**67.0%**
	Positive	71.0%	71.0%	71.0%
	Negative	53.3%	53.3%	53.3%

The results show that the performance for the negative category is reduced. Each classifier alone is better than the voting scheme because of discrepancies among classifiers. For example, if two classifiers vote for a positive polarity, the word will be classified as positive, even if the SVM classifies it correctly as negative.

The resulting lexicon has 30,773 affective words, including different meanings for each word. For example, the number accompanying the word *rosa* (3) indicates that the third meaning was used to determine its polarity.

In the resulting dictionary, we included the polarity, the word, and the gloss that describes the sense of the word. For example, in Table 2, we show some results of the dictionary. The first column indicates the polarity as positive (p) or negative (n), the second column indicates the classified word (*hurgar* / delve), and the last column indicates the meaning of the word.

Table 2. Excerpt from the affective Spanish dictionary

Results		
Polarity	**Word**	**Gloss**
n	***hurgar***	*remover una cosa, escarbar.*
n	***chorrear***	*caer un líquido a chorro.*
n	***roer***	*raspar con los dientes una cosa, generalmente un alimento, arrancando parte de ella.*
n	***rosa3***	*mancha rosácea que sale en el cuerpo.*
n	***aberrar***	*andar errante, equivocarse, aberración, aberrante.*
p	***comer2***	*tomar la comida principal del día [en mi casa comemos a las dos].*
p	***contribuir***	*pagar las contribuciones o impuestos.*

With respect to analysis of errors, we identified some errors when analysing the classified words. For example, if the word sense is related to a common animal (error type 1), the classified word has no positive or negative polarity at all; human annotators also hesitate how to classify these sorts of words. In the second error type, the definition is very short at word level; it is difficult that classifiers assign the correct label due to lack of context.

Table 3. Errors in the classification process

Error	Example
1. common animal	*p|perro|nombre común, mamífero carnívoro, doméstico* (p|dog| common animal, carnivorous mammal, domestic)
2. short definition	*n|rivera|arroyo* (n|stream|creek) *n|revuelto|de revolver* (n|mess-up|to mess up)

5. Conclusions

In this paper, we have presented a method that generates a large affective lexicon using supervised learning techniques for Spanish. We evaluated the results obtained using a test set with 3,000 words that were selected randomly and labeled by hand. The training set used consisted of 5,222 words from two affective lexicons (SEL and PL). The resulting lexicon has 30,773 words, classified as positive or negative words, including different meanings for each word. The precision obtained was 67.0%. It shows that the quality of our lexicon outperforms that of lexicons whose entries are classified using just one classifier.

In the future, we will not use stems to train the classifiers. In order to improve their performance, we will use lemmatisers, part of speech taggers, or syntactic n-grams (Sidorov et al., 2014) for example, since they have already been used for this purpose with good results.

References

Agarwal, A., Biadsy, F., & Mckeown, K. (2009). Contextual phrase-level polarity analysis using lexical affect scoring and syntactic n-grams. In *Proceedings of EACL 2009* (pp. 24-32). Greece: ACL. Retrieved from http://dx.doi.org/10.3115/1609067.1609069

Banea, C., Mihalcea, R., & Wiebe, J. (2011). Multilingual sentiment and subjectivity analysis. In D. Bikel & I. Zitouni (Eds.), *Multilingual natural language processing applications: from theory to practice.* Boston: IBM Press. Retrieved from http://people.cs.pitt.edu/~wiebe/pubs/papers/multilingualSubjBookChap2011.pdf?

Díaz-Rangel, I., Sidorov, G., & Suárez-Guerra, S. (2014). Creación y evaluación de un diccionario marcado con emociones y ponderado para el español. *Onomazein, 29*, 1-23. Retrieved from http://dx.doi.org/10.7764/onomazein.29.5

Esuli, A., & Sebastiani, F. (2006). SentiWordNet: a publicly available lexical resource for opinion mining. In *Proceedings of LREC 2006*, (pp. 417-422). Italy.

Gelbukh, A., Sidorov, G., & Ledo-Mezquita, Y. (2003). On similarity of word senses in explanatory dictionaries. *International Journal of Translation*, *15*(2), 51-60.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: an update. *ACM SIGKDD Explorations Newsletter, 11*(1), 10-18. Retrieved from http://dx.doi.org/10.1145/1656274.1656278

Miller, G. A. (1995). WordNet: a lexical database for English. *Communications of the ACM, 38*(11), 39-41. Retrieved from http://dx.doi.org/10.1145/219717.219748

Navigli, R. (2009). Word sense disambiguation: a survey. *ACM Computing Surveys (CSUR), 41*(2). Retrieved from http://dx.doi.org/10.1145/1459352.1459355

Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical Methods in Natural Language Processing* (pp. 79-86). USA: ACL. Retrieved from http://dx.doi.org/10.3115/1118693.1118704

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval, 2*(1-2), 1-135. Retrieved from http://dx.doi.org/10.1561/1500000011

Pérez-Rosas, V., Banea, C., & Mihalcea, R. (2012). Learning sentiment lexicons in Spanish. In *Proceedings of LREC 2012*, (pp. 3077-3081). Turkey: ELRA.

Porter, M. F. (2006). Spanish stemming algorithm. Retrieved from http://snowball.tartarus.org/algorithms/spanish/stemmer.html

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing and Management*, *24*(5), 513-523. Retrieved from http://dx.doi.org/10.1016/0306-4573(88)90021-0

Saralegi, X., & San Vicente, I. (2013). Elhuyar at TASS 2013. In *Proceedings of XXIX Congreso de la Sociedad Española de Procesamiento de lenguaje natural*. *Workshop on Sentiment Analysis at SEPLN (TASS2013)* (pp. 143-150). Madrid: SEPLN.

Sidorov, G., Miranda-Jiménez, S., Viveros-Jiménez, F., Gelbukh, A., Castro-Sánchez, N., Velásquez, F., Díaz-Rangel, I., Suárez-Guerra, S., Treviño, A., & Gordon, J. (2012). Empirical study of machine learning based approach for opinion mining in tweets. In *Advances in Artificial Intelligence (MICAI 2012)* (pp. 1-14). Mexico: Springer.

Sidorov, G., Velasquez, F., Stamatatos, E., Gelbukh, A., & Chanona-Hernández, L. (2014). Syntactic N-grams as machine learning features for natural language processing. *Expert Systems with Applications*, *41*(3), 853-860. Retrieved from http://dx.doi.org/10.1016/j.eswa.2013.08.015

Wiebe, J., & Riloff, E. (2005). Creating subjective and objective sentence classifiers from unannotated texts. In *Proceeding of CICLing-05, International Conference on Intelligent Text Processing and Computational* (pp. 486-497). Mexico: Springer. Retrieved from http://dx.doi.org/10.1007/978-3-540-30586-6_53

Published by Research-publishing.net, not-for-profit association
Dublin, Ireland; Voillans, France, info@research-publishing.net

New perspectives on teaching and working with languages in the digital era
Edited by Antonio Pareja-Lora, Cristina Calle-Martínez, Pilar Rodríguez-Arancón

Disclaimer: Research-publishing.net does not take any responsibility for the content of the pages written by the authors of this book. The authors have recognised that the work described was not published before, or that it was not under consideration for publication elsewhere. While the information in this book are believed to be true and accurate on the date of its going to press, neither the editorial team, nor the publisher can accept any legal responsibility for any errors or omissions that may be made. The publisher makes no warranty, expressed or implied, with respect to the material contained herein. While Research-publishing.net is committed to publishing works of integrity, the words are the authors' alone.

Trademark notice: product or corporate names may be trademarks or registered trademarks, and are used only for identification and explanation without intent to infringe.

Copyrighted material: every effort has been made by the editorial team to trace copyright holders and to obtain their permission for the use of copyrighted material in this book. In the event of errors or omissions, please notify the publisher of any corrections that will need to be incorporated in future editions of this book.

Typeset by Research-publishing.net
Cover design and frog picture by © Raphaël Savina (raphael@savina.net)

ISBN13: 978-1-908416-34-6 (Paperback - Print on demand, black and white)
Print on demand technology is a high-quality, innovative and ecological printing method, with which the book is never 'out of stock' or 'out of print'.

ISBN13: 978-1-908416-35-3 (Ebook, PDF, colour)
ISBN13: 978-1-908416-36-0 (Ebook, EPUB, colour)

Legal deposit, Ireland: The National Library of Ireland, The Library of Trinity College, The Library of the University of Limerick, The Library of Dublin City University, The Library of NUI Cork, The Library of NUI Maynooth, The Library of University College Dublin, The Library of NUI Galway.

Legal deposit, United Kingdom: The British Library.
British Library Cataloguing-in-Publication Data.
A cataloguing record for this book is available from the British Library.

Legal deposit, France: Bibliothèque Nationale de France - Dépôt légal: mai 2016.