ARTICLE IN PRESS

Information Processing and Management xxx (2014) xxx–xxx

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

# A Spanish semantic orientation approach to domain adaptation for polarity classification

M. Dolores Molina-González, Eugenio Martínez-Cámara *, M. Teresa Martín-Valdivia, L. Alfonso Ureña-López

*SINAI Research Group, University of Jaén, Campus Las Lagunillas, E-23071 Jaén, Spain*

ARTICLE INFO

*Article history:*
Received 14 August 2013
Received in revised form 3 October 2014
Accepted 5 October 2014
Available online xxxx

*Keywords:*
Spanish opinion mining
Sentiment lexicon
Domain adaptation

ABSTRACT

One of the problems of opinion mining is the domain adaptation of the sentiment classifiers. There are several approaches to tackling this problem. One of these is the integration of a list of opinion bearing words for the specific domain. This paper presents the generation of several resources for domain adaptation to polarity detection. On the other hand, the lack of resources in languages different from English has orientated our work towards developing sentiment lexicons for polarity classifiers in Spanish. The results show the validity of the new sentiment lexicons, which can be used as part of a polarity classifier.



## 1. Introduction

Opinion Mining (OM) is defined as the computational treatment of opinion, sentiment, and subjectivity in text. This new area of research is becoming more and more important mainly due to the growth of social media where users con- tinually post contents on the web in the form of comments, opinions, emotions, etc. The OM discipline combines Natural Language Processing (NLP) with data mining techniques and includes a large number of tasks (Pang & Lee, 2008). One of the most widely studied tasks is the polarity classification of reviews. This task focuses on determining the overall sentiment-orientation (positive or negative) of the opinions contained within a given document.

Although different approaches have been applied to polarity classification, the mainstream basically consists of two major methodologies. On the one hand, the Machine Learning (ML) approach (also known as the supervised approach) is based on using a collection of data to train the classifiers (Pang, Lee, & Vaithyanathan, 2002). On the other hand, the approach based on Semantic Orientation (SO) does not need prior training, but takes into account the positive or negative orientation of words (Turney et al., 2002). This method, also known as the unsupervised approach, makes use of lexical resources like lists of opinion words, lexicons, dictionaries, etc. Both methodologies have their advantages and drawbacks. For example, the ML approach depends on the availability of labelled data sets (training data), which in many cases are impossible or difficult to achieve. On the contrary, the SO strategy requires a large amount of linguistic resources which generally depend on the language, and often this approach obtains lower recall because it depends on the presence of the words comprising the lexicon in the document in order to determine the orientation of opinion. In this paper we focus on the generation of linguistic resources to tackle the problem of polarity classification using an unsupervised approach.

* Corresponding author.
*E-mail addresses:* mdmolina@ujaen.es (M. Dolores Molina-González), emcamara@ujaen.es (E. Martínez-Cámara), maite@ujaen.es (M. Teresa Martín-Valdivia), laurena@ujaen.es (L. Alfonso Ureña-López).

http://dx.doi.org/10.1016/j.ipm.2014.10.002
0306-4573/

While opinions and comments on the Internet are expressed in any language, most research in OM is focused on English texts. However, languages such as Chinese, Spanish or Arabic, are even more present on the web. Thus it is important to develop resources to help researchers to work with these languages. The work presented herein is mainly motivated by the need to develop polarity classification systems and resources in languages other than English. Specifically, in this paper we deal with Spanish reviews. We present an experimental study over the SFU Review Corpus[1] (Brooke, Tofiloski, & Taboada, 2009), which is a comparable corpus that includes opinions of several topics in English and in Spanish in different domains.

One of the open problems in OM is that of domain adaptation. Although movie reviews have been the most studied domain in sentiment analysis, a wide range of areas are being investigated such as political debates, hotels or music. However, when we train a classifier using a specific domain we need to adapt it in order to apply it to another domain. For example, the sentence "Definitively, you should read the book" most likely refers to positive polarity for Book reviews but negative sentiment for Movie reviews.

Thus the problem of domain adaptation is attracting more and more attention in OM. In this paper we carry out an experimental study of domain adaptation of linguistic resources for Spanish reviews in different domains. We have used the Spanish version of SFU, which includes 400 reviews for 8 different domains. We have generated several lists of opinionated words integrating knowledge from the different domains and we have compared the results obtained. A corpus-based approach is followed with the aim of adapting a general-purpose sentiment lexicon to a specific domain by integrating lists of opinion bearing words. iSOL[2] (Molina-González, Martínez-Cámara, Martín-Valdivia, & Perea-Ortega, 2013) is the general-purpose sentiment lexicon chosen. The Spanish version of the SFU corpus was the corpus selected for the adaptation process due mainly to the fact that it covers 8 different domains. Following different heuristics, which will be described later, the most frequent opinion bearing words are appended to iSOL. Several experiments were carried out with the goal of assessing the new domain-specific sentiment lexicons. The analysis of the results shows the validity of the new lists.

The paper is organised as follows: Section 2 briefly describes other papers that study non-English sentiment polarity classification and, specifically work related to Spanish OM. In addition, we include some papers studying the domain adaptation problem. In Section 3 we explain the different resources used. Sections 4 and 5 present the experiments carried out and discusses the main results obtained. Finally, we outline conclusions and further work.

## 2. Related work

In this study we focus on two open problems in opinion mining: non-English polarity classification and the domain adaptation problem. Next, we will comment on some papers that have inspired our work.

### 2.1. Non-English polarity classification

There are some interesting papers that have studied the problem using non-English collections. For example, Denecke (2008) worked on German comments collected from Amazon. These reviews were translated into English using standard machine translation software. Then the translated reviews were classified as positive or negative, using three different classifiers: LingPipe3, SentiWordNet (Esuli & Sebastiani, 2006) with classification rule, and SentiWordNet with machine learning. In (Zhang, Zeng, Li, Wang, & Zuo, 2009) Chinese sentiment analysis is applied on two datasets. In the first one euthanasia reviews were collected from different web sites, while the second dataset was about six product categories collected from Amazon (Chinese reviews). Ghorbel and Jacot (2011) used a corpus with movie reviews in French. They applied a supervised classification combined with SentiWordNet in order to determinate the polarity of the reviews. In (Agić, Ljubešić, & Tadić, 2010) a manually annotated corpus is presented with news on the financial market in Croatia. In (Rushdi-Saleh, Martín-Valdivia, Ureña López, & Perea-Ortega, 2011) a corpus of movies reviews in Arabic annotated with polarity was presented and several experiments using machine learning techniques were performed.

Regarding Spanish, there are also some interesting studies. For example, Banea, Mihalcea, Wiebe, and Hassan (2008) proposed several approaches to cross lingual subjectivity analysis by directly applying the translations of opinion corpus in English to training an opinion classifier in Romanian and Spanish. This study showed that automatic translation is a viable alternative for the construction of resources and tools for subjectivity analysis in a new target language. In (Brooke et al., 2009) several experiments dealing with Spanish and English resources are presented. They conclude that although the ML techniques can provide a good baseline performance, it is necessary to integrate language-specific knowledge and resources in order to achieve an improvement. They proposed three approaches: the first one uses Spanish resources generated manually and automatically. The second one applies ML to a Spanish corpus. The last one translates the Spanish corpus into English and then applies the SO-CAL (Semantic Orientation CALculator), a tool developed by themselves (Taboada, Brooke, Tofiloski, Voll, & Stede, 2011). Cruz, Troyano, Enriquez, and Ortega (2008) manually recollected the MuchoCine (MC) corpus in order to develop a sentiment polarity classifier based on semantic orientation. The corpus contains annotated Spanish movie reviews from the MuchoCine website.[3] The MC corpus was also used in (Martínez-Cámara,

[1] urlhttp://www.sfu.ca/œmtaboada/research/SFU_Review_Corpus.html.
[2] The iSOL resource is freely available for research purpose at urlhttp://sinai.ujaen.es/?p=1202.
[3] urlhttp://www.muchocine.net/.

Please cite this article in press as: Dolores Molina-González, M., et al. A Spanish semantic orientation approach to domain adaptation for polarity classification. *Information Processing and Management* (2014), http://dx.doi.org/10.1016/j.ipm.2014.10.002