Noname manuscript No.
(will be inserted by the editor)

Spanish Corpora for Sentiment Analysis: a Survey

María Navas-Loro · Víctor Rodríguez-Doncel

Received: date / Accepted: date

Abstract Corpora play an important role when training machine learning systems for sentiment analysis. However, Spanish is underrepresented in these corpora, as most primarily include English texts. This paper describes 20 Spanish-language text corpora – collected to support different tasks related to sentiment analysis, ranging from polarity to emotion categorization. We present a brand-new framework for the characterization of corpora. This includes a number of features to help analyze resources at both corpus level and document level. This survey – besides depicting the overall landscape of corpora in Spanish – supports sentiment analysis practitioners with the task of selecting the most suitable resources.

Keywords sentiment analysis · corpora · opinion mining · polarity · emotion

1 Introduction

Corpora are generally understood as large collections of digital texts, where every document has been marked as belonging to one or more specific categories or in which some documents and/or fragments have been tagged with additional information. Corpora are necessary to train statistical systems which

This work has been partially funded by a Predoctoral grant from the I+D+i program of the Universidad Politécnica de Madrid and by project Datos 4.0 (TIN2016-78011-C4-2-R).

María Navas-Loro (Corresponding author)
Ontology Engineering Group, Universidad Politécnica de Madrid, Spain
Tel.: +34-913363670
E-mail: mnavas@fi.upm.es

Víctor Rodríguez-Doncel
Ontology Engineering Group, Universidad Politécnica de Madrid, Spain
Tel.: +34-913363672
E-mail: vrodriguez@fi.upm.es

will be used for sentiment analysis. However, complete well-documented corpora in Spanish are scarce, despite widespread acceptance that having a variety of high-quality resources is critical to be able to achieve good results.

The recent study on 'Language equality in the digital age' [36], commissioned by the Directorate-General for Parliamentary Research Services of the European Parliament, reported that the Spanish language is underrepresented on the web and as regards the number of language resources available.

The objective of this paper is to collect and systematically describe the Spanish-language corpora available for sentiment analysis – across different sectors, formats and even classifications, with categories ranging from basic polarity to complex emotion annotation. This review considers different aspects that could be useful for a number of NLP tasks, and provides a short comment for each of the corpora and details the tasks for which they are most suitable. To the best of our knowledge, no analysis of this kind exists for Spanish-language corpora in the field of sentiment analysis.

The paper is organized as follows: Section 2 presents an evaluation framework we have developed specifically for corpora which could be used for sentiment analysis. The framework makes the systematic study and comparison of corpora easier, and identifies a number of features that can be of help to characterize them. In Section 3, we analyze the main features of each corpus individually. Section 4 presents a comparison of the corpora, in relation to the different aspects identified by the evaluation framework. Finally Section 5 presents some conclusions derived from the review and some future lines of research in Spanish-language sentiment analysis.

2 Evaluation framework

The goal of this section is to identify the key features of the corpora available for sentiment analysis. Evaluations of quality, e.g. whether annotations are reliable or diverse enough, have been excluded from this framework.

Corpora for sentiment analysis can be found scattered across the web and also cataloged in language resource repositories. Table 1 identifies some of the most popular repositories along with a selection of the metadata fields used to describe the resources. Nevertheless, these repositories gather corpora from different domains and even media types (such as video or audio); therefore, they only consider very generic characteristics (such as size) but not the specific features relevant for sentiment analysis. Although each repository relies on different metadata schemas, this paper will use the most commonly employed. We also add useful practical information (as a relation to other resources) specific to sentiment analysis. In Table 1, *Ling.* stands for linguality in terms of monolinguality or multilinguality, *Lang.* is the language of the resource, *Avail.* describes the availability of the resource, *Lic.* refers to license information, *Size* describes the amount of data (in any format, like text or tokens), *Text format* refers to the kind of file and representation of the data and *Source* deals with information about where the original data was collected. Other information

Table 1 Information available in different online repositories and catalogues.

	Ling.	Lang.	Avail.	Lic.	Size	Use / Domain	Text Format	Source	Others
LRE MAP[1]	X	X	X	X	X	X	-	-	Conference info, paper, Production Status
META-SHARE[2]	X	X	X	X	X	X	X	-	Creation, version, funding project, documentation...
LDC[3]	-	X	X	X	FT	X	FT	X	Release date, IDs, documentation, citation, funding project...
ReTeLe[4]	X	X	X	X	X	X	X	FT	IDs, funding project...
ELRA[5]	X	X	price	-	X	FT	-	FT	IDs, funding project

found in the different repositories is shown in the last column. *FT* stands for free text and means that the information was not found as metadata but can be included as part of a free text description.

The corpus features have been categorized into two groups in this evaluative framework: features describing the corpus (*corpus metadata*) and features describing the corpus entries (*corpus data*). The second group has, in turn, been broken down into further categories.

2.1 Corpus metadata features

The features that we will use to describe the corpora for sentiment analysis include the following:

– **Topic and source** describes the domain and the documents' source. The topic can be general, a concrete sector or a number of sectors, whereas sources can be as diverse as web forums, opinion websites or social networks. This is a significant characteristic, since the source usually determines constraints and specific expressions, and document characteristics which need to be taken into account when using them for sentiment analysis. Twitter posts, for instance, have their own peculiarities – such as

[1] `http://www.elra.info/en/catalogues/lre-map/`. All the references have been visited in November 2018.

[2] `http://www.meta-share.org`

[3] `https://catalog.ldc.upenn.edu`

[4] `http://catalogo.retele.linkeddata.es`

[5] `http://catalog.elra.info`

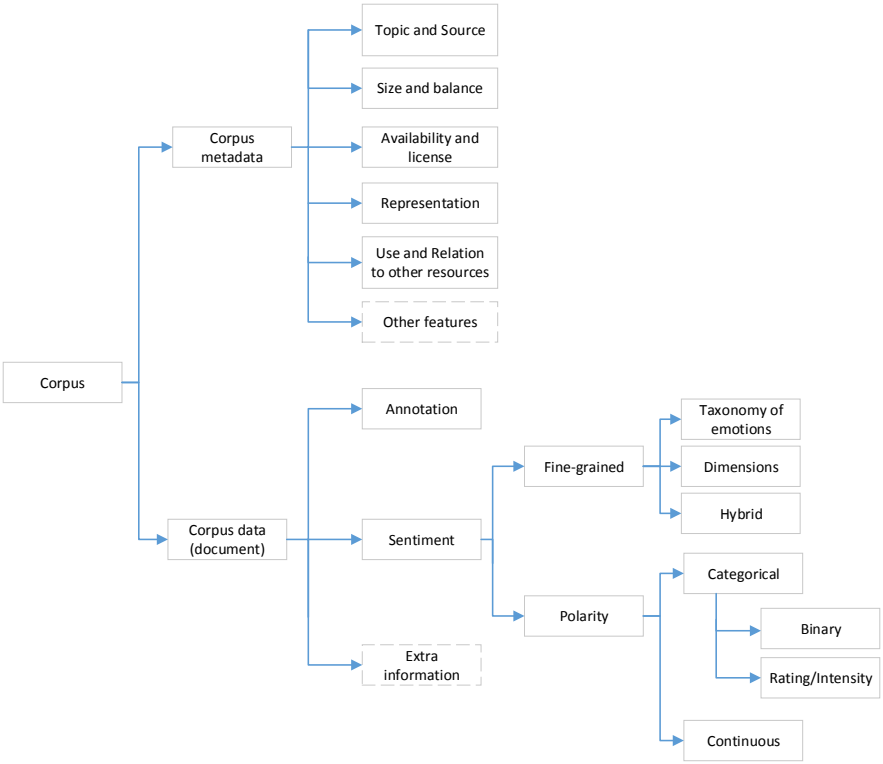

Fig. 1 Framework for evaluation: taxonomy of features.

hashtags and restricted character numbers – and require different analytical methodology compared to a long review or blog post. Social networks also tend to impose restrictions derived from their privacy and/or copyright policies.

- **Size and balance.** The number of entries in the corpus, as well as entry length (usually as average wordcount). Other aspects related to size include minimum entry size, and the balance of categories — unbalanced corpora are harder to use for certain tasks.
- **Availability and license.** Whether the resource can be accessed directly, and any restrictions on permitted use. The resource might be fully available online, just partially available or not available at all. Access to the resources is sometimes straightforward, sometimes requires registration, and sometimes the researchers may need to be contacted. Finally, we analyze the resources' licenses.
- **Representation** considers the format of the corpus entries, from plain text to structured XML or RDF.

- **Use and relation to other resources** describes the purpose for which the corpus was originally collected and the current use to which it is put. Additionally, some corpora are linked to other resources such as lexicons or lexical databases (e.g. Wordnet).
- **Other features.** This captures other information which is particularly relevant for specific corpora, such as details about who provided the opinion or what the opinion's target was. This category also includes information about inter-annotator agreements and data collection methodology – where available.

2.2 Corpus data features

Some features of the evaluation framework describe aspects at corpus entry level. Three kinds of features can be of interest:

- **Sentiment** describes how the opinion, sentiment or appraisal is represented. Sometimes fine-grained sentiments are used at corpus-entry level (such as hate and love), whereas other times only a simple polarity value is given.
  - **Fine-grained sentiment.** The latest efforts in sentiment analysis try to go beyond polarity by finding the specific emotions present in an opinion. These emotions can be represented in several ways, from fixed schemes to dimensional values, and even hybrid approaches.
    - *Taxonomy of emotions.* Several classifications of human emotions have been proposed in the literature, especially in the field of Psychology, such as those of Ekman [11], Plutchik [32] and Tomkins [46]. Additionally, works like Nakamura's dictionary [26] and Shaver's taxonomy [44] provide a basis for emotion classification; dedicated taxonomies have sometimes been built for concrete domain applications – usually by combining or modifying previous work. Choice of emotion categorization usually focuses on the most extended systems, and tries to make the resources as reusable as possible; that is, capable of being integrated with previous work in the area. There have also been efforts to map particular systems onto other classification schemas. Finally, another major aspect which influences corpus choice is cultural background, and mainly language, since emotions and subjectivity are intrinsically linked to social, cultural and educational factors.
    - *Dimensions.* The dimensional approach to emotion classification has received a lot of attention recently. These have been developed since Russell's initial approaches [42] with the creation of the Circumplex Model of Affect – and the definition of one of the most widely-used dimensional vectors, PAD (Pleasure, Arousal, Dominance) [24] – some proposals have evolved from previous ideas (such as Fontaine's addition of "unpredictability" to the PAD vector [12]

and the Hourglass model proposed in Sentic, based on Plutchick's work [7]); others have been been built more or less from scratch, such as Cochrane's eight dimensions for emotions [8]. Contributions from other fields such as Medicine have also gained traction. This inlcudes Lövheim's Cube of Emotions [18] which relates three monoamine neurotransmitters to the basic emotions outlined in Tomkins' Affect Theory [46].
    - *Hybrid.* Corpora can also be tagged with different categories, or define some kind of mapping onto other classification options.
- **Polarity** is the most used and most basic categorization for sentiment analysis, often only specifying whether an opinion is positive or negative. Nevertheless, phenomena such as irony, polysemy, language style or lack of context notably increase not only classification difficulties, but also the task of actually annotating a corpus. As with emotions, ratings can be either categorical or continuous.
    - *Categorical*, with this approach each document in the corpus can belong to one of several predefined categories, which may be binary (positive or negative, but sometimes including categories for neutral texts and/or texts with no emotions) or in the form of a rating. Discrete ratings can take the form of an integer or simply a text qualifier, such as '*awful*', '*bad*' or '*very good*'.
    - *Continuous.* A decimal number in a predefined range.
- **Annotation.** This includes additional linguistic- or sentiment-analysis related information, such as POS tags, emoticons or hashtags.
- **Extra information.** Some corpora include information about other aspects which can be relevant for sentiment analysis, such as the presence of irony, features like hashtags or mentions, degree of subjectivity or information about the author of each text.

3 Spanish Language Corpora

This section briefly describes different corpora found in the literature or on the web and summarizes their main distinctive features. The methodology we followed consisted of the following steps: (i) defining an evaluation framework; (ii) web search of resources, literature examination and language catalogue browsing; (iii) retrieval and verification of the resources and study of the related academic publication (if existing) and (iv) comparative analysis of the corpora and refinement of the evaluation framework defined in (i).

3.1 HOpinion

The corpus HOpinion [37] contains 17,934 opinions (2,388,848 words) extracted from *TripAdvisor*[6] (dating from 2004 to 2011), mainly evaluating

[6] `https://www.tripadvisor.es/`

different accommodation options. Opinions are given about hotels, hostels, apartments and other resorts, and also about tourist attractions, restaurants and other points of interest. Ratings are expressed as a number of stars between one and five, but are not balanced within the corpus. HOpinion contains 841 one-star reviews, 1,269 two-star reviews, 3,468 three-star reviews, 6,244 four-star reviews and 6,112 five-star reviews – positive reviews predominate. The corpus is available as both plain text and as a database, and it includes information about the following aspects:

- The target of the opinions (named *items* in the corpus), with free-text description, type, category and location.
- The users who wrote the opinions, usually just user name and place of origin, and – if available – age, gender and self-description of travel style, timing, company and preferences.
- The reviews, with the user and item it relates to, date of the post and a free-text comment together with a 5-star rating, which tends to confirm the sentiment present in the free-text part of the opinion.

Part of the corpus (4,740 texts) contains additional annotations for lemmas and morphology following the EAGLES label schema[7]. To gain access to the corpus, an account must be created and verified[8], although the resource is also available in other locations[9]. Published under a CC-BY 3.0 license, no connections to other resources are provided, besides the TripAdvisor URLs of the places evaluated.

This corpus, or fragments thereof [48,49], has been used for testing opinion-mining algorithms [13] and also for automatic register classification tasks as it includes morphological annotation and linguistic style description [38].

3.2 COAR

COAR[10] is a corpus built from *TripAdvisor* which contains 2,202 opinions (retrieved from December 2012 to January 2015) about hotels and restaurants, rated from 1 to 5 stars; concretely, there are 565 opinions with one star, 246 with two, 188 with three, 333 with four and 870 rated as five-star. The corpus is reasonably balanced, with 1,203 positive opinions and 811 negative. Fields in the dataset include the rating (one to five), name of the place, review title (which tends to be rich in sentiment), review text, and comment date. It is freely available as an XLSX file encoded in UTF-8. No license is publicly specified. Name, organization and e-mail address are required to obtain the resource.

[7] `http://blade10.cs.upc.edu/freeling-old/doc/tagsets/tagset-es.html`

[8] `http://clic.ub.edu/corpus/es/node/106`

[9] `http://johnrbto.com/hopinion/`

[10] `http://sinai.ujaen.es/coar/`

3.3 COAH

COAH [25] is a corpus built from TripAdvisor with 1,816 opinions (239,749 words) on hotels and restaurants – ranked from 1 to 5 stars. Opinion distribution is as follows: 312 are 1-star opinions, 199 2-star, 285 3-star, 489 4-star and 531 5-star (with an unbalanced ratio of 1020 negative opinions to 511 positive). This corpus provides linguistic statistics about the corpus, such as the number of unique words (154,297), tokens (272,446) and lemmas (239,749) and the average sentence length (23.25 words). The corpus is available as a UTF-8 XML file, which includes each review's score, title (called *abstract*) and the review text itself. The focus of the corpus is hotels, all of which are Andalusian, ten from each province (five well-rated and five poorly-rated), with more than 20 opinions each. It can be retrieved by providing name, organization and contact e-mail – no licensing information is publicly specified.

3.4 COST

COST [23] is a corpus of pre-processed informal posts from *Twitter* obtained via queries. These search queries (from the 3rd and 4th of March 2011) were based on a set of predefined positive (such as `:)`, `:D` and `XD`) and negative (like `:(`, `:-/` or `D:`) emoticons. Those containing elements from both polarities were deleted. The corpus was pre-processed, removing newlines, URLs, and repeated letters; laugh expressions were also normalized. The result is a balanced corpus containing 34,634 tweets, 17,317 positive and 17,317 negative[11]. The corpus is available as a UTF-8 CSV file upon request via email[12]. The corpus contains information on date of retrieval, the user who posted, the text itself and a polarity value. This value was derived automatically from the query which harvested the tweets: if a tweet was found because of positive emoticons, polarity was set to '1'; if retrieved because of the negative emoticons, polarity was set to '0'. The corpus creators carried out a number of different experiments on the corpus, and concluded SVM was the best option among the sentiment analysis classification algorithms they tested.

3.5 COPOS

COPOS [31] (Corpus Of Patient Opinions in Spanish) is a corpus of opinions about healthcare, collected on 3rd December 2015. It is composed of patients opinions about the medical attention they received extracted from the website *Masquemedicos*[13]. The corpus contains 743 reviews (2,009 sentences and 32,365 words) about 34 medical specialties; each review includes the user name,

[11] These are the numbers in the corpus description; however, the data itself varied slightly: 34,615 tweets (17,311 negative and 17,304 positive).

[12] `http://sinai.ujaen.es/cost/`

[13] `http://masquemedicos.com`

a negative and a positive text[14], location, date, specialty, name of the doctor evaluated and polarity – expressed from 0 to 5 stars. The authors considered as negative reviews rated up to 2 stars and positive from 3 to 5. The corpus is very unbalanced: 3 reviews have 0 stars, 88 1 star, 18 2 stars, 35 3 stars, 51 4 stars and 548 have 5 stars. The corpus can be obtained via email[15] as a UTF-8 CSV file. It has been used in conjunction with a Dutch medical opinion corpus, COPOD [15].

3.6 Trip-MAML

Trip-MAML [53] is an extension in both Spanish and Italian of a previous English opinion dataset extracted from TripAdvisor [19], following the same annotation protocol; the dataset is freely available online[16]. The corpus comprises about 500 TripAdvisor reviews from the ten most visited cities in Spain, annotated at a sentence-level with different aspects (*Sleep Quality, Value, Building, Service, Rooms, Cleanness, Food, Location, Other* and *Not Related*) with three levels of polarity: *positive, neutral/mixed* and *negative*. 1 to 5 star ratings are also provided for each aspect as well as an overall rating. The dataset is provided as several JSON files, divided by annotator or by train/test, and also includes some linguistic information. Zafra provides more information on the annotation process and inter-annotator agreement, along with some baseline experiments on the corpus [53].

3.7 DOS

The Drug Opinion Spanish (DOS) corpus [16], was built from comments on the website *mimedicamento*[17] which contains patients' opinions about and experiences with different medications.

The corpus contains 877 reviews of the 30 most-commented upon medications as of 14 March 2017. Each of these comments has an id (*rid*) following the pattern `nameOfTheDrug(component)__numberOfReview`. The corpus does not include information about the writer (such as age and/or gender) nor whole comment star-ratings for different categories (such as general, efficiency or ease of ingestion); this is despite these characteristics being available on the website. The dataset has, on the other hand, been annotated at an aspect-based level regarding side-effects, with each of the 887 texts having been split into sentences (3,784 in total and with ids the same as the respective *rid* plus `:numberOfTheSentenceInTheText`), which contain *opinions* about side-effects. Each *opinion* contains a category (in the version we analyzed this was always "side-effect"), a target (containing the textual reference to this

[14]The source website asks for most and least positive aspects of users' experiences

[15]`http://sinai.ujaen.es/copos-2/`

[16]`http://hlt.isti.cnr.it/trip-maml/`

[17]`https://www.mimedicamento.es`

side-effect), offset in the sentence, polarity (*positive*, *negative* or *netrual*) and intensity (*high, medium, low*).

As regards corpus balance, there are 2,230 mentions of side-effects: 98 positive, 2,119 negative and 13 neutral. Each side-effect mentioned is classified as high-intensity (655), medium-intensity (1,486) or low-intensity (89). The number of side effects does not appear to be related to the length nor the number of sentences of a review; this is because a single sentence can express several opinions, while a long text may have just one. The corpus is freely available[18] (after providing name, institution and email) as a single XML file with UTF-8 encoding.

3.8 SemEval-2016 ABSA Spanish Dataset

Task 5 of the SemEval international workshop[19] (Semantic Evaluation) provided datasets in different languages for Aspect Based Sentiment Analysis (ABSA), including one in Spanish. The Spanish dataset [33] has two sets of annotations on the same corpus of 913 restaurant reviews (sourced from two different Spanish restaurant-review websites[20]) matching two different subtasks:

- Subtask 1: Sentence-level Aspect-Based Sentiment Analysis, with annotations in the form of tuples of aspect categories (such as *price* or *quality* for entities like *food* or *restaurant*), opinion target expressions and sentiment polarity (*positive*, *negative* and *neutral*). The set of aspects to analyze are limited and defined in Appendix A of the task reporting document [33].
- Subtask 2: Text-level Aspect-Based Sentiment Analysis, with global polarity annotations on aspects of each review.

As reported in the task description, the training dataset contains 627 texts and 2,070 sentences; the test dataset, 286 texts and 881 sentences. The annotation produced 2,720 tuples for the first subtask and 2,121 tuples for the second in training, and 1,072 and 881 tuples in the test sample. The corpus is very unbalanced: the training set – used for the first subtask – includes 1,925 positive sentences out of 2,070. Links to download different parts of the corpus (via Meta-Share) are available on the website[21]. The corpus was released as a collection of UTF-8 XML files (following a tree structure similar to the one in the DOS corpus described in Section 3.7). The part containing Subtask 1 is divided into sentences where the presence of polarity is given as offset indications – with the category and target to which it related. The part with the Subtask 2 has a list of different opinions on the whole text for each review, without offset and target – just polarity and categories. More information about the

[18] `http://sinai.ujaen.es/dos/`

[19] `http://alt.qcri.org/semeval2016/`

[20] `http://www.bcnrestaurantes.com/`, `http://www.restaurantes-zaragoza.es`

[21] `http://alt.qcri.org/semeval2016/task5/index.php?id=data-and-tools`

annotation process and the evaluation results is available online, as well as related resources in other languages [33].

All eight parts of the dataset (four for each targeted subtask: one for training and three for test, marked as either Phase A, Phase B or Gold annotations) are accessible from Task 5 of the SemEval website along with corpora in other languages, and via a MetaShare repository[22]. It is restricted to non-commercial academic uses. Besides its use in the SemEval Task 5, this corpus has been employed in further studies, such as Vázquez et al.'s polarity algorithm [47].

3.9 EmotiBlog

EmotiBlog [3] is a corpus which was obtained by collecting Spanish, English and Italian blog posts on three different topics. The dataset has been annotated at document, sentence and word-level following the "Emotiblog" annotation scheme, which tries to capture linguistic elements that imply subjectivity. Information about these elements, such as verbs or anaphora, and annotation for each is also available [3]. Additional information on polarity (*positive*, *negative*) and degree of polarity (*high, medium* or *low*), emotions (using their own taxonomy), modifiers and confidence levels about these annotations is also included in the "Emotiblog" annotation format, along with a number of other phenomena.

The corpus's creators adopted several ideas from the MPQA corpus [52], one of the most widely used corpora in English sentiment analysis, including the topics (e.g. the Kyoto Protocol, the Spanish corpus's topic). There are 100 different texts (the link to the original blog is provided for each), but not all are annotated. The corpus is licensed for research purposes only, and can be downloaded[23] as individual text files (one per opinion), some with their respective annotated XML GATE file (with different encodings) [10], or as a single XML file.

3.10 Spanish Movie Reviews

The Spanish Movie Review corpus [9] consists of 3,878 film reviews extracted from the MuchoCine website[24]. These reviews had to fulfill a series of requirements about the amount and the quality of the opinions available and the licensing of the content, among others.

Each review is provided as an ANSI encoded XML file containing the summary of the review and its body, as well as the author's user-name, the review title and the rating information. Files containing linguistic information (including morphemes, lexemes and Wordnet synsets, among others) and dependency trees generated by the external tool FreeLing [2] are also provided – for the

[22] `http://metashare.ilsp.gr:8080/repository/search/?q=semeval+2016+Spanish`

[23] `https://gplsi.dlsi.ua.es/gplsi13/es/node/344`

[24] `http://www.muchocine.net`

summary and opinion body separately. Additionally, the original reviews are easily retrievable from the web using the URI pattern `http://www.muchocine.net/critica/XXXX`, where `XXXX` is the number of the review.

As regards balance, 351 reviews have been rated as 1-star, 923 as 2-stars, 1,253 as 3-stars, 890 as 4-stars and 461 as 5-stars, adding up therefore to a fairly well-balanced corpus. Besides the tests reported by the authors in the main paper, further studies have also made use of this resource [21,22, 49]. The dataset is publicly available[25] under a CC BY attribute license; an English version of the corpus has also been created by automatic translation [21].

3.11 SFU Spanish Review Corpus

The Simon Fraser University (SFU) Spanish Review corpus [5] was developed as a Spanish parallel to the English SFU Review Corpus [45]. It contains 400 reviews on different topics: cars, hotels, books, cell phones, music, computers and movies. This version includes washing machines rather than the cookware-themed reviews in the English corpus. The opinions were extracted from the website *Ciao*[26], where users rate various kinds of products from one to five stars. Each review is stored in the corpus as free text in an ANSI text file in the format `POL_NSTARS_NUM.txt`, where `POL` represents polarity (`yes` for positive, 4 or 5 stars in the original review, and `no` for negative, 1 or 2 stars), `NSTARS` is the number of stars rated by the reviewer and `NUM` is the number of the review within its polarity-target subset. Reviews are balanced, and each category includes 50 reviews (25 for each polarity and similar amounts for each rating).

This corpus has been used in Vilares et al.'s syntatic study [49], among others. The corpus can be downloaded[27] following registration. Its annotations were extended in the SFU Review-NEG corpus, described below.

3.12 SFU Review-NEG

The SFU Review-NEG [20] is an extended version of the SFU Spanish Review Corpus (see previous Subsection 3.11) which adds annotations on negations and their scope. Distinct from the original, the files are provided as .TBF.XML (but maintaining the same naming convention), and include new linguistic annotations (tokens, lemmas, and POS) and have been divided into sentences. Although the corpus is more focused on dealing with negation than on sentiment analysis, it is a valuable resource since it complements the original corpus with POS annotations, tokens, lemmas and events. Apart from its focus on negation, the influence of how negation affects polarity is annotated

[25] `http://www.lsi.us.es/~fermin/corpusCine.zip`

[26] `http://www.ciao.es/`

[27] `http://www.sfu.ca/~mtaboada/download/downloadCorpusSpa.html`

considering *polarity_modifiers* (*reduction* and *increment*). It provides analysis as to whether there is a *change* in the polarity due to negation structures – as well as other similar information. The polarity category is included at sentence level.

A detailed study of the validity of this corpus is also available, with an extensive explanation on the annotation format [17]. This extra annotation on negation has been used, for instance, by Periñán-Pascual [30]. The corpus is available for research purposes only[28].

3.13 RepLab 2013 Dataset

The RepLab competitive evaluation exercise for Online Reputation Management (RepLab) released a dataset of tweets in English and Spanish for its 2013 edition. It includes, among others, "Polarity for reputation classification" tags (*positive*, *negative*, *neutral*) which should identify if a tweet has potential positive or negative impacts on a company's reputation. This method presents a number of differences from standard sentiment analysis, as explained by Amigó et al. [1]; these tags refer to the reputational impact, which may differ from the sentiment in the opinion (for instance, *'being sorry about someone's death'* implies negative emotions but is positive for the entity's reputation). Therefore, when the tweet is unrelated to the entity's reputation, no polarity is provided.

The dataset includes tweets that mentioned entities from four different areas (Automotive, Banking, University and Music) on Twitter. It is divided into test (96,848) and train (45,689) subsets, adding up 142,527 tweets (63,442 positive, 30,493 neutral and 16,415 negative), out of which only 28,983 are in Spanish. The dataset is provided as several UTF-8 .dat files, including information such as the *tweet id*, the *topic* and the *entity*, but not the text (although links to tweets are also provided, these might have been deleted and therefore irretrievable). Additionally, the content of the links referred to in the tweets and the entities' Wikipedia pages are also available.

The corpus can be used for test purposes via the EvAll service[29] and downloaded for free[30]. As with the other datasets and in order to respect copyright, tweets are merely referenced by their identifier. The corpus user is then expected to run a script to retrieve the actual content.

3.14 EmIroGeFb

EmIroGeFb [34] is a corpus of comments on three different topics (politics, football and celebrities) from *Facebook* labeled by three different taggers with

[28] `http://sinai.ujaen.es/sfu-review-sp-neg/`, `http://clic.ub.edu/corpus/es/node/171`

[29] `http://www.evall.uned.es/`

[30] `http://nlp.uned.es/replab2013/`

the emotions described in the text, the presence of irony and the gender of the author. For each of the three topics, four different official Facebook pages in Spanish for each entity were selected (four political parties, four football teams and four celebrities – two TV celebrities, one actor and one singer), and 400 comments were extracted for each topic, 200 from men and 200 from women.

The corpus is provided as a single UTF-8 XML file. The following information is provided for each comment:

- Its *Facebook ID* (so the text can be retrieved from the website following the instructions provided; it should be noted that not all comments are still online, so not all the texts listed in the corpus are still available).
- The user's *gender*.
- The *topic* it was matched to.
- Three different `annotatorX` sets (with `X` the reference for the annotator). Each annotator tagged the texts with the following:
  - The six Ekman's emotions (`joy, surprise, sadness, anger, disgust` and `fear`), one tag for each – marked *true* or *false*, and an additional tag `no-emotion` set to *true* if none of Ekman's emotions were tagged by the annotator. Several emotions can be tagged as *true* in the same comment.
  - The tag `irony` is set as *true* if the annotator considered that there was irony in the comment.

The inter-annotator agreement and statistics on the corpus are provided, along with analysis by the corpus creators. This includes the percentage of comments per emotion, which is unbalanced: *surprise* is present in 32.50% of the comments and *joy* in 28.17%, whereas *sadness* was tagged in only 6.33% and *fear* in 0.25%. This corpus can be downloaded (without the comments' text but with the IDs) for research and social analysis purposes[31]. It has been used for author profiling in the main [41] and is one of the few Spanish corpora which deals with irony and emotions.

3.15 SAB/MAS Corpus

The Spanish Corpus for Sentiment Analysis towards Brands (SAB Corpus) [29] is a collection of 4,546 tweets which mentioned brands from different sectors (food, automotive, banking, beverages, sports, retail and telecoms). These tweets are tagged using the publicly available guidelines for implementing a taxonomy of four different emotions (*love, satisfaction, happiness* and *trust*) and their direct opposites (*hate, dissatisfaction, sadness* and *fear*). As the name implies, the corpus was built especially for marketing purposes. The polarity and the brand which evoked the emotions (since several brands can appear in the same text) are included in each tweet, along with its Twitter ID (as the text itself cannot be published). Information about the inter-annotator agreement and statistics on the corpus are available on the same website.

[31] `http://ow.ly/uQWEs`

The corpus is published as RDF linked data and includes mappings to external databases like Thomson Reuters' PermID[32] and DBpedia[33] for each brand. Several ontologies have been reused for representation, such as Marl and Onyx [43] for Sentiment and Emotions, SIOC [4] for post-representation and GoodRelations [14] for marketing. A dedicated vocabulary was also developed for this corpus; both the corpus and the vocabulary are downloadable[34].

The annotations in the SAB corpus have recently been extended and curated, forking into the MAS corpus (Marketing Analysis in Spanish) [28]. First, repeated or extremely similar tweets were erased, resulting in a final dataset of 3,763 tweets. Second, two new marketing-related dimensions were added, the Marketing Mix (the target of the opinion, including the categories *product, price, promotion* and *place*); and the Purchase Funnel (the stage in the purchase journey where the opinion's author was, between *awareness, evaluation, purchase, post-purchase* and *ambiguous*). The creators also provided new statistics on different linguistic aspects of the corpus. This information can be found on its website[35]; the corpus itself is available from Zenodo[36], and both versions are downloadable under a CC-BY 4.0 license.

3.16 TASS General corpus*

The TASS General corpus was released for first edition of the TASS challenge in 2012 [51], and was used by all subsequent editions up until 2017. It contains 68,017 tweets retrieved from 154 well-known celebrities from different countries and covers ten topics (politics, entertainment, economy, music, soccer, films, technology, sports, literature and others). It is provided as train (7,219) and test (60,798) sets, including just the *tweet ID*, the *user ID*, the *creation date* and the *topics* of each tweet. Each message has been semi-automatically tagged with polarity at entity level (when existing, also tagging the entity referred to) and with global polarity; five different levels of polarity are considered (`P+, P, NEU, N, N+`), along with a no sentiment tag – `NONE`. There is no information available about the overall balance of the corpus, but the level of agreement (as *agreement* or *disagreement*) is provided.

Various parts of the corpus can be downloaded as UTF-8 XML files from the websites of the different TASS editions, after signing a License Agree-

[32] `https://permid.org/`

[33] `http://dbpedia.org/`

[34] `http://sabcorpus.linkeddata.es/`

[35] `http://mascorpus.linkeddata.es/`

[36] `https://zenodo.org/record/1293493#.W3O_V-gzbIU`

*An annual Spanish workshop, TASS for sentiment analysis, releases datasets for different tasks every year, some of them newly built for that year and others reused from previous editions. The tagged datasets come from different editions of this workshop. In order to be granted access to TASS corpora, a Research/Non-Commercial License Agreement must be signed and sent to the organizers; more information can be found at the website of each edition (where the schema for reading the XML files is also provided).

ment[37]. The latest, 2017, edition also provides a gold standard for QREL format[38] and text in the test XML file. The 2016 version can be tested online via the EvAll platform.

3.17 TASS Politics corpus*

The TASS Politics corpus was used in the 2013 TASS workshop [50]. It contains 2,500 tweets that mention the four main Spanish political parties from the 2011 Spanish General Elections (*PP, PSOE, IU* and *UPyD*). Unlike other TASS corpora, only three levels of polarity were considered (`P, NEU, N` plus `NONE`), and tags were added at both entity-level and tweet-level. The format is the same as in the aforementioned TASS General corpus described in Section 3.16, but an extra `source` attribute for the political party has been added to the entity.

The original dataset is available as a collection of UTF-8 XML files on the TASS 2013 website and that of subsequent editions[39].

3.18 TASS Social-TV corpus*

The Social-TV corpus, released for TASS 2014 [40] and also used at TASS 2015 [6], contains 2,773 tweets written during a 2014 football match, divided into train (1,773) and test (1,000) and developed for aspect-based sentiment analysis tasks. Polarity is tagged at three levels (`P, NEU` and `N`, with no distinction made between `NEU` and `NONE`). The polarity tags refer to particular aspects of the match (concrete, such as names of *teams, coaches* or *players*; or more abstract entities like *fans* and *broadcasting*). For this corpus, text is provided for each tweet (also identified by its *tweet ID*), as annotations for the *aspect* and the polarity are marked in the text itself. The dataset is available as UTF-8 XML and QREL from these editions' websites.

3.19 TASS STOMPOL corpus*

The STOMPOL corpus (Spanish Tweets for Opinion Mining at aspect level about POLitics) [6] is comprised of 1,284 tweets (784 in the training set and 500 for test) related to political topics (*Economics, Health System, Education, Political Party* and *Other aspects*) related to the main political parties (*PP, PSOE, IU, Podemos, Cs* and *UPyD*). The data was collected during the Spanish general election campaign in 2015. Two annotators (and a third in the case of a disagreement) tagged polarity at three levels (`P, NEU, N`) for each

[37] `http://www.sepln.org/workshops/tass/tass_data/download.php`

[38] `http://www.sepln.org/workshops/tass/2017/`

[39] `http://www.sepln.org/workshops/tass/2013/corpus.php`

writer's opinion about each aspect and the political parties referred to; however, the tweet's overall polarity was excluded. As with the Social-TV corpus (Section 3.18), aspect annotation was done in-text, and includes `sentiment` tags which encompass attributes like the tagged *aspect*, the *entity* (political parties) referred to and the *polarity*.

The corpus has been released as UTF-8 XML (and the gold standard, as QREL) and is available on the websites of the editions which used it[40].

3.20 TASS InterTASS corpus*

The InterTASS corpus (International TASS Corpus) was created for the 2017 edition of the TASS workshop, and includes 3,413 tweets divided into train (1,008), development (506) and test (1,899). The corpus was released for a tweet-level sentiment analysis task. Each tweet's polarity (within the range `P, NEU, N, NONE`) was tagged by three different annotators, and also includes the *tweetid*, *creation date* and *user ID*. As regards balance, there are between $30\% - 33\%$ positive tweets, $40\% - 43\%$ negative, $11\% - 13\%$ neutral and $12\% - 14\%$ tweets which present no sentiment, depending on the particular subset.

As with other TASS corpora, the different parts of the dataset are available from the website – in this case the 2017 edition – as XML and QREL (the gold standard).

4 Results and Corpora Comparison

Table 2 summarizes the main features of each corpus, using the evaluation framework presented in Section 2.

4.1 Topic and Source

The different topics covered by each corpus are presented in Table 3, as well as the target(s) of those opinions (such as organizations, brands or persons, among others). While some corpora cover a number of different separated sectors or topics (such as the SAB corpus and SFU corpora), others focus on a specific area (such as restaurants or politics) and some are just general – as a result of having used temporal or text filtering criteria for collection; this is the case, for instance, with the COST corpus, which was harvested from Twitter by searching for polarized emoticons.

Sometimes the topics have been identified by using the specific target of the opinion, such as organizations (presented as O in Table 3) (sports teams (O_T), brands (O_B), political parties (O_P), universities (O_U) or people (P).

[40] `http://www.sepln.org/workshops/tass/2015/tass2015.php#corpus`, `http://www.sepln.org/workshops/tass/2016/tass2016.php#corpus`

Table 2 Spanish Corpora available for sentiment analysis. Columns correspond to the aspects considered in Fig. 1

ID	Corpus Features				Document Features	
Corpus	Topic *Source*	Size *Balance*	Availability *License*	Represent. *Text*	Sentiment Analysis	Aspects & Other info
HOPINION	Tourism *TripAdvisor*	17,934 *No*	Registration *CC-BY 3.0*	mySQL,txt *Yes*	Rating (1-5)	-
COAR	Restaurants *TripAdvisor*	2,202 *No*	Registration -	xslx *Yes*	Rating (1-5)	-
COAH [25]	Tourism *TripAdvisor*	1,816 *No*	Registration -	xml *Yes*	Rating (1-5)	-
COST [23]	General *Twitter*	34,634 *Yes*	On request -	csv *Yes*	Polarity (0/1)	-
COPOS [31]	Patient opinions *MasQueMedicos*	743 *No*	On request -	csv *Yes*	Rating (0-5)	-
Trip-MALM [53]	Tourism *TripAdvisor*	500 *No*	Free -	json *Yes*	Polarity (positive, negative, neutral)	Aspects
DOS [16]	Opinions on drugs *MiMedicamento*	877 *Yes*	Registration -	xml *Yes*	Polarity (positive, negative, neutral)	Intensity
SemEval-2016 ABSA Spanish Dataset [33]	Restaurants *Restaurant websites*	913 -	Registration *MS NC NoRED*	xml *Yes*	Polarity (positive, negative, neutral)	Category, target
EmotiBlog [3]	Kyoto Protocol *Blogs*	100 -	Free -	txt,xml *Yes*	Emotions, polarity (positive, negative)	Intensity and more
Spanish Movie Reviews [9]	Cinema *MuchoCine*	3,878 *Yes*	Free *CC-BY 2.1 ES*	xml *Yes*	Rating (1-5)	-
SFU Spanish Review Corpus [5]	Several items *Ciao*	400 *Yes*	Registration *GNU GPL v3*	xml *Yes*	Rating (1-2,4-5)	-
SFU Spanish NEG Review Corpus [20]	Several items *Ciao*	400 *Yes*	Registration *CC-BY 1.0*	xml *Yes*	Rating (1-2,4-5)	Interaction negation-polarity and more
RepLab 2013 Dataset [1]	Automotive, Music, Banking, University *Twitter*	28,983 *No*	Free -	dat *No (IDs)*	Polarity of reputation (positive, negative and neutral)	Entities
EmIroGeFb [34]	Politics, Football, Celebrities *Facebook*	1,200 -	Free -	xml *No (IDs)*	Ekman emotions	Gender, topic, presence of irony
SAB Corpus [29]	Brands from several sectors *Twitter*	4,846 *No*	Free *CC-BY 4.0*	n3 *No (IDs)*	Emotions, polarity	Brands mentioned
TASS General [51]	Personalities *Twitter*	68,017 *No*	On Request *Own*	xml,qrel *No (IDs)*	Polarity (P+,P,NEU, N,N+,NONE)	Level of agreement
TASS Politics [50]	Politics *Twitter*	2,500 *No*	On Request *Own*	xml *No (IDs)*	Polarity (P,NEU, N,NONE)	Level of agreement, entity
TASS Social-TV [40]	Sports *Twitter*	2,773 *No*	On Request *Own*	xml,qrel *Yes*	Polarity (P,NEU,N)	Aspects
TASS STOMPOL [6]	Politics *Twitter*	1,284 *No*	On Request *Own*	xml,qrel *Yes*	Polarity (P,NEU,N)	Aspects
TASS InterTASS	General *Twitter*	3,413 *No*	On Request *Own*	xml,qrel *Yes*	Polarity (P,NEU, N,NONE)	Aspects

Table 3 Topics covered by the corpora analyzed.

Topic	Restaurants	Hotels	Medical	Food	Automotive	Banking	Beverages	Sports	Retail	Telecom	Books	Music	Computers	Cinema	Washing Machines	Universities	Politics	TV	General
HOPINION		X																	
COAR	X	X																	
COAH	X	X																	
COST																			X
COPOS			Doc																
Trip-MAML		X																	
DOS			Drug																
SemEval'16	X																		
EmotiBlog																	KP		
Spanish Movie Reviews														X					
SFU & SFU-NEG		X			X					CP	X	X	X	X	X				
RepLab'13					O_B	O_B						P				O_U			
EmIroGeFb								O_T				P		P			O	P	
SAB				O_B	O_B	O_B	O_B	O_B	O_B	$O_{B_{pr}}$									
General																			X
Politics																	O		
Social-TV																		P	
STOMPOL																	O		
InterTASS																			X

On other occasions, corpora cover very specific subtopics in a domain, such as opinions about doctors (Doc) and medication (Drugs) in the medical domain or Cell Phones (CP) in telecoms or the Kyoto Protocol (KP) in politics; these have been underlined in Table 3.

Most generalized Spanish corpora without a specific domain focus have been built from Twitter content (Table 4), whereas domain-specific corpora have been created from specialized sites. Thus, TripAdvisor is used to collect opinions about the tourism sector (targets like hotels, sightseeing and restaurants); Ciao, as a source of purchasing comments; and websites like MasQueMedicos, MiMedicamento and MuchoCine provide opinions about doctors, medications and films. In addition to availability, several factors related to the source should be taken into account when selecting a corpus.

Intrinsic characteristics of the text format. While comments on opinion forums tend to present a more complex text structure (with titles, introduction, body, and close) and unlimited length (and more curated and refined text, especially in the case of blogs); tweets are length-limited, and usually include abbreviations and special expressions (such as hashtags), orthographic errors, typos and word omission – phenomena which also appear on other social networks such as Facebook. This category also elucidates how polarity differs depending on the source: for social networks the opinion has to be manually annotated based on the free text alone, usually involving a degree of subjective judgment and potential misunderstandings due to lack of context and details about the comment author's experience and intentions; in opinion forums the free text is reinforced by some predefined ranking system that leaves no doubt about the polarity the user wanted to express. Additionally, sources may have dissemination restrictions, as is generally the case with social networks; these platforms do not usually allow the text or user name to be published in public corpora.

Table 4 Sources datasets were collected from. From left to right, `Twitter` *(Tw)*, `Facebook` *(FB)*, `TripAdvisor` *(TripAdv)*, `MasQueMedicos` *(MasQueMed)*, `MiMedicamento` *(MiMed)*, `MuchoCine`, `Ciao`, the websites `bcnrestaurantes` *(BCNR)* and `restaurantes-zaragoza` *(RZar)* and different blogs.

Type	Social Networks		Opinion Forums						Other
Source	Tw	FB	TripAdv	MasQueMed	MiMed	MuchoCine	Ciao	BCNR & RZar	Blogs
HOPINION			X						
COAR			X						
COAH			X						
COST	X								
COPOS				X					
Trip-MAML			X						
DOS					X				
SemEval'16								X	
EmotiBlog									X
Spanish Movie Reviews						X			
SFU							X		
SFU-NEG							X		
RepLab'13	X								
EmIroGeFb		X							
SAB	X								
General	X								
Politics	X								
Social-TV	X								
STOMPOL	X								
InterTASS	X								

Collection method. It is not only the kind of source which is important, but also how it was obtained: the Facebook comments used by EmIroGeFb, for instance, were obtained from various entities' Facebook pages, as a consequence the opinions found there were clearly targeted at the entity. This also happens with opinion forums, where the target (whether a movie, a hotel or a medication) is clearly defined in the opinion itself. With Twitter, however, just mentioning a brand or product does not always imply a clear opinion.

Intention of the user. This is also valuable information: specialized forums, focused on reviews of concrete topics, usually receive more grounded opinions – probably based on previous experience and generally considering a variety of aspects – while Social Network opinions (especially Twitter) tend to be more vague.

Finally, the specific domain should, obviously, also be taken into account when selecting a source, as opinion forums tend to have very specific targets or at least some way to filter them and adapt to the needs of the user. The benefits and drawbacks associated with each of these characteristics should be considered when choosing a dataset for a particular task.

4.2 Size and balance

Corpora present disparate different sizes and distribution forms, as illustrated by Fig. 2 and 3, and Tables 5 and 6.

Focusing on size, despite the fact that copora containing over 15,000 text documents are available (such as HOPINION, COST, RepLab 2013 and TASS General), most corpora contain less than 4,000 documents, some less than 1,000. This is especially relevant for machine learning purposes. Exact figures can be found in Fig. 2, and although the graph's y-axis is logarithmic due to the sizeable gap between the largest corpora and the smaller corpora, it does clearly show the differences and size rankings.

As regards balance, most corpora are unbalanced with respect to polarity and star-ratings (as shown in Fig 3), which are the most common options for sentiment analysis categorization; there are almost always more positive reviews than negative or neutral. This also happens in the datasets which consider emotion: texts expressing no emotion are more common, and moderate or positive emotions in reviews such as *satisfaction, dissatisfaction, joy* and *surprise* are more common than negative or extreme emotions like *hate, anger* and *disgust*. Unfortunately, no information about potential overlaps is provided by any of the emotion-annotated corpora. Some of the taxonomies used to annotate the documents include emotions which recur across the various different datasets (see Section 4.7).

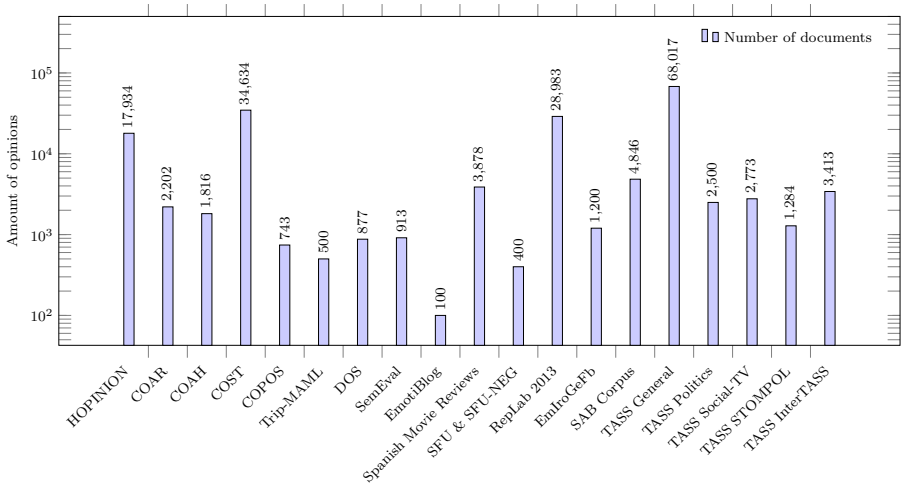

Fig. 2 Number of text documents per corpus. Please note that for sake of visibility and due to the big gap between the most populated corpus (68,000) and the less populated corpora (c. 100 texts), the y-axis is logarithmic.

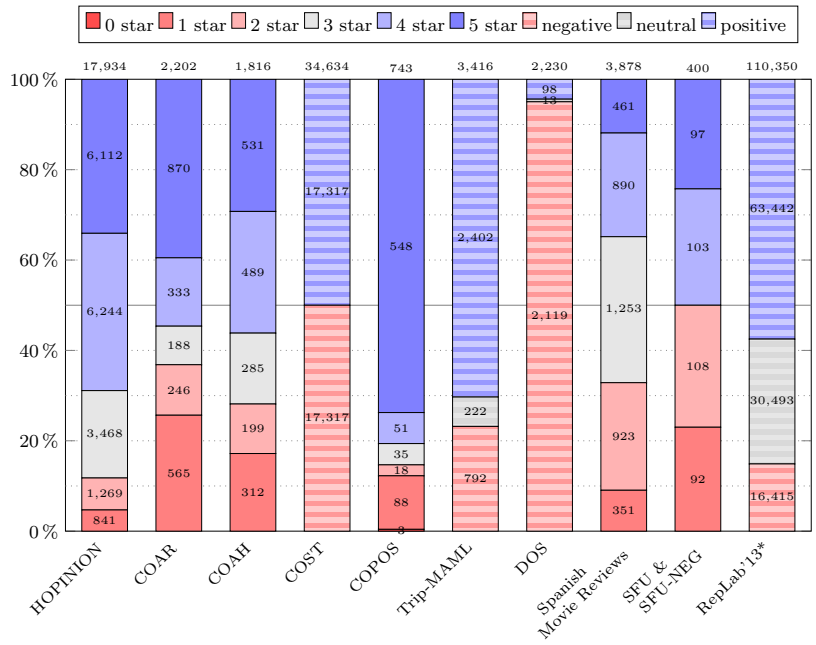

Fig. 3 Balance of star-rated and polarized corpora per opinion when available (e.g. SemEval just provides polarity per sentence). Trip-MAML numbers are given at sentence-level, while DOS provides polarity on side effects. The asterisk denotes that RepLab'13 figures are for the whole corpus, since no available information on just the Spanish part is provided.

Table 5 Number of text documents for each polarity. Some corpora provide both star-rating and polarity (such as SFU and SFU-NEG). For corpora without balance information, *X* indicates the type of annotations provided. The asterisk has the same meaning as in Fig 3.

	Polarity						Stars						Total
	N+	N neg	NEU neutral	P pos	P+	NONE	0	1	2	3	4	5	Amount
HOPINION	-	-	-	-		-	-	841	1,269	3,468	6,244	6,112	17,934
COAR	-	-	-	-	-	-	-	565	246	188	333	870	2,202
COAH	-	-		-	-	-	-	312	199	285	489	531	1,816
COST	-	17,317	-	17,317	-	-	-	-	-	-	-	-	34,634
COPOS	-	-	-	-	-	-	3	88	18	35	51	548	743
Trip-MAML	-	792	222	2,402	-	-	-	X	X	X	X	X	500
DOS	-	2,119	13	98	-	-	-	-	-	-	-	-	877
SemEval'16	-	X	X	X	-	-	-	-	-	-	-	-	913
EmotiBlog	-	X	-	X	-	-	-	-	-	-	-	-	100
Spanish Movie Reviews	-	-	-	-	-	-	-	351	923	1,253	890	461	3,878
SFU & SFU-NEG	-	200	-	200	-	-	-	92	108	-	103	97	400
RepLab'13	-	16,415*	30,493*	63,442*	-	-	-	-	-	-	-	-	28,983*
General	X	X	X	X	X	X	-	-	-	-	-	-	68,017
Politics	-	X	X	X	-	X	-	-	-	-	-	-	2,500
Social-TV	-	X	X	X	-	-	-	-	-	-	-	-	2,773
STOMPOL	-	X	X	X	-	-	-	-	-	-	-	-	1,284
InterTASS	-	X	X	X	-	X	-	-	-	-	-	-	3,413

Table 6 Number of emotions in corpora, where *happ* stands for *happiness*, *sat* for *satisfaction*, *dissat* for *dissatisfaction* and *surp* for *surprise*. Emotiblog contains more emotions than those shown in the table, and no information about the distribution in terms of the overall balance of the corpus.

	Emotions												Others		Total
	love	hate	happ	sadness	trust	fear	sat	dissat	joy	surp	anger	disgust	None	More	Amount
EmotiBlog	-	-	X	-	-	X	-	-	X	X	X	X	X	X	100
EmIroGeFB	-	-	-	76	-	3	-	-	338	390	151	129	262	-	400
SAB	207	127	178	48	605	117	759	576	-	-	-	-	3,142	-	4,546

4.3 Availability and License

Documents in corpora which can be used for sentiment analysis are protected by copyright laws, and their availability online does not grant corpus users any rights besides mere access to the resource. In particular, creating derivative works (further annotation, for example) or redistribution is not permitted unless specific permission has been granted. Many resources, however, are published along with a license which specifies the precise terms and conditions under which they can be used. These terms sometimes prevent use in commercial applications, restrict use to academic environments or merely require the acknowledgement of the original creators. These conditions should be published in a clear, structured manner [39]; however, corpus publishers generally fail to specify any licensing information, thus hampering their use in commercial set-ups.

The selection, annotation and arrangement of documents in a corpus can eventually also be protected by copyright, so the actual chain of permissions required to lawfully redistribute a corpus would then include the document creators' permission plus the corpus creators'. As a consequence, some corpora's annotations are publicly available on the web (for example the sentiment of tweets), but not the documents themselves (tweets are only referenced, as the corpus publisher is not authorized to redistribute them). As a result, using the corpus requires the download of information from both the corpus creators' website and the document publisher's.

Table 7 illustrates how Spanish corpora for sentiment analysis are licensed. From left to right, the type of license (if available) *(Type)*, whether the resource can be freely used or not *(Free)*[41], or just for research purposes *(Academic)*, whether the resource is freely available *(Free)*, whether the user needs to identify themselves (name, affiliation etc.) to download it *(Id)*, create an account *(Account)*, directly request it from someone *(Request)* or submit a License Agreement to someone in order to retrieve it *(LA)*, and if the text is available in the corpus or if it has to be retrieved via ID, for example.

Access to the corpora analyzed is generally straightforward, as most of the documents are available online or accessible upon registration.

[41] It is considered to be freely available if not explicitly mentioned to be otherwise.

Table 7 Licenses, rights of reuse, type of access and availability of text for each corpora.

	License	Rights of reuse		Access					Text	
	Type	Free	Academic	Free	Id	Account	Request	LA	No	Yes
HOPINION	CC-BY 3.0	X				X				X
COAR	-	X			X					X
COAH	-	X			X					X
COST	-	X					X			X
COPOS	-	X					X			X
Trip-MAML	-	X		X						X
DOS	-	X			X					X
SemEval'16	MS NC NoReD		X			X				X
EmotiBlog	-		X	X						X
Spanish Movie Reviews	CC-BY 2.1 ES	X		X						X
SFU	GNU GPL v3	X			X					X
SFU-NEG	CC-BY 1.0		X		X					X
RepLab'13	-		X	X					ID	
EmIroGeFb	-	X		X					ID	
SAB	CC-BY 4.0	X		X					ID	
General	Own		X					X	ID	
Politics	Own		X					X	ID	
Social-TV	Own		X					X		X
STOMPOL	Own		X					X		X
InterTASS	Own		X					X		X

Table 8 Representation format.

Representation	mySQL	txt	xlsx	xml	csv	json	dat	n3	qrel
HOPINION	X	X							
COAR			X						
COAH				X					
COST					X				
COPOS					X				
Trip-MAML						X			
DOS				X					
SemEval'16				X					
EmotiBlog		X		X					
Spanish Movie Reviews				X					
SFU				X					
SFU-NEG				X					
RepLab'13							X		
EmIroGeFb				X					
SAB								X	
General				X					X
Politics				X					
Social-TV				X					X
STOMPOL				X					X
InterTASS				X					X

4.4 Representation

Information about the file format of each corpus is shown in Table 8.

4.5 Use and relation to other resources

Describing the provenance of corpus documents is a key practice for reusability and reproducibility of research procedures. SINAI corpora (including COAH, COAR, COPOS, COST, DOS or SFU-Review NEG corpora) use a representation schema (*xml*) which includes provenance – as does the TASS corpora. Besides isolated uses in the literature – already reported in the individual analysis of each corpus in the previous section – the source of the corpus is also related to its use. TASS corpora have all been used in the TASS workshop, which publishes each year's results and whose participants need to submit a paper explaining their approaches. These resources serve well as reference material for new algorithms, and their results provide good baseline for comparison. The same applies to SemEval'16, where Spanish datasets – and other language resources following similar criteria – are available and have been tackled by different teams across the world. Other kinds of relations between the various resources, besides their provenance, follow.

- The COPUS corpus has been developed alongside a similar Dutch corpus of opinions about medical issues, the COPOD corpus.
- The SemEval 2016 corpus is one of the datasets used in the SemEval competition. Similar datasets are also available for other languages (English, Arabic, Chinese, Dutch, French, Russian and Turkish).
- Emotiblog was based on the English MPQA corpus, and uses the Emotiblog annotation schema.
- The Spanish Movie Reviews corpus links the words in the reviews to their synset in Wordnet.
- The SFU corpus (and therefore its continuation SFU Review-NEG) was based on the English SFU Review Corpus, built from the same source and containing almost the same topics.
- The RepLab 2013 dataset links the target entities in the tweets to their Wikipedia entries.
- The SAB/MAS corpora link each target brand in the tweets to their Thomson Reuters' PermID database and their DBPedia resource.

4.6 Annotation

Information on linguistic and sentiment related annotation is provided in Table 9 for each corpus. Most corpora do not offer this kind of information, and when they do, they simply provide the output of an automatic tool like FreeLing.

4.7 Sentiment features

Table 5 and 6 quantitatively describe ratings information. As noted in Section 4.2, most sentiment analysis annotations are based on polarity categorizations,

Table 9 Additional linguistic or sentiment analysis-related annotation provided by some corpora. In some cases, more annotation is provided, including some which follow their own annotation scheme (e.g. Emotiblog).

	Morpho	Lemmas	Intensity	Subjectivity	WordNet	Dependency	Negation	Others / own scheme
HOPINION	X	X						
Trip-MAML	X	X						
DOS			X					
EmotiBlog			X	X				X
Spanish Movie Reviews	X	X			X	X		
SFU-NEG							X	X

both binary and/or star-rated. A relation between binary and star-based ratings can be made, as is done explicitly by the SAB corpus and the SFU/SFU-NEG corpora, according to the following schema:

- 1 and 2 starred opinions => negative ; 4 and 5 starred opinions => positive 3 starred opinions => neutral.
- 1 and 2 starred opinions => negative ; 3, 4 and 5 starred opinions => positive
- 1 starred opinions => N+ ; 2 starred opinions => N; 3 starred opinions => NEU / NONE & NEU ; 4 starred opinions => P; 5 starred opinions => P+

Alternative mapping options for the polarity annotations used in TASS corpora for the SAB corpus's emotion categorization have been described previously [27]. With respect to emotion taxonomy mappings, some nominal overlaps between the three corpora analyzed do exist (as shown in Table 6); however, this also affects secondary emotions not noted in the table or the corpus. Emotion taxonomies tend to consist of some primary emotions used for annotation and some secondary emotions which can be considered as pertaining to primary emotions, whether also annotated or not. These aid annotation tasks. This is the case of the emotion *joy* in EmIroGeFB, which includes *happiness*, and *happiness* is at the same time a main emotion in the SAB corpus. Producing mappings between these emotions is therefore straightforward in most cases, with the exception of non-polarized emotions such as *surprise*. A list of secondary emotions is usually provided with the corpus or in its annotation guidelines. Emotiblog is remarkable in this context, since it includes more than 60 emotions clustered into 15 groups – which could also be considered emotions in and of themselves.

Given the complexity of some emotion taxonomies [3], the comparison of sentiments in Table 6 should be considered with caution: the same term for an emotion can be differently defined and convey a broader, narrower or even different meaning depending on the specific corpus.

The last aspect to be considered relates to the scope of the annotation. Most corpora provide text-level annotation, but some include a richer description

that details both aspects and targets of the sentiment – the ABSA SemEval'16 corpus is notable for providing a number of different sentiments at entity level.

5 Conclusions and Future work

There are a number of corpora for sentiment analysis in Spanish, and each can serve a different purpose well. Their most prominent features can be summarized as follows:

- **HOPINION** provides opinions on hotels with additional information about the users, such as age and place of origin when available, and also linguistic information. It is the only corpus that is also available as a database, which makes it easy to create queries to retrieve specific parts of the corpus.
- **COAR** is an easy to read corpus (in *xlsx*) which can be used in conjunction with the COAH corpus and other corpora from TripAdvisor.
- **COAH** provides extensive linguistic data, such as tokens, senses and sentence length.
- **COST** would fit a task focused on emoticons, since its tweets were collected and classified according to their very presence. COST provides both the text and IDs of the messages.
- **COPOS** is the only corpus that provides opinions about doctors and is related to a Dutch corpus on the same topic.
- **Trip-MAML** is a recent a multilingual corpus with evaluation in several languages of new aspects such as *Sleep* (sleep quality) in the tourism sector. An analysis of the inter-annotator agreement across the different languages is also included.
- **DOS** provides extensive sentence-level information on medication side-effects, such as polarity and intensity towards a target, and as far as we know, it is the only effort of this kind in Spanish.
- **SemEval 2016 ABSA corpus** uses both text and sentence-level annotations, and since it is part of an international evaluation, several approaches for the Spanish corpus and its equivalent corpora in other languages are provided; it is, therefore, well-suited to multilingual sentiment analysis tasks.
- **Emotiblog** might be an option if interest lies in subjectivity, word-level annotation or just need to work with the Emotiblog schema; it is also multilingual, with texts also in English and Italian.
- The **Spanish Movie Reviews** corpus provides linguistic information (morphemes, lexemes, morphosyntactic and associated probability), sentence dependency trees and WordNet synsets.
- The **SFU corpus** includes several topics with balanced opinions from the well-known product review site Ciao; it provides the actual text and both polarities and star ratings. It has been forked into the **SFU Review-NEG** corpus, with additional linguistic annotations (POS annotations, lemmas etc.), and there is also an analogous English version. The corpus's particular

focus is negation and its interaction with polarity, making it a unique and valuable resource in Spanish.
- The **RepLab 2013** dataset is a multilingual corpus of English and Spanish tweets. It is focused on polarity with in relation to the reputation of an entity; there is also a link to the relevant Wikipedia article of each entity. It can also be tested online as a preloaded benchmark via EvAll.
- **EmIroGeFb** is the only corpus that provides gender and irony information. It is annotated with Ekman's well-known emotion taxonomy.
- The **SAB corpus** might be the best option if both polarity and emotions related to marketing purposes are required. It is published as linked data and contains references to external resources, and the annotations of a proportion of the tweets has been later extended into the **MAS corpus**.
- **TASS corpora** are among the largest Spanish language corpora online, cover a broad range of topics and are widely used by the Spanish academic community. In addition, they are updated regularly and serve as a test-bed for the algorithms which compete in a yearly challenge.

This article presents the results of a manual survey of Spanish corpora for sentiment analysis, and has contributed an evaluation framework which can be referenced in the self-description of future sentiment-analysis corpora.

Dataset annotation is now going beyond polarity, and other aspects such as irony are becoming a major subject of study [35]. As annotations become more and more complex and yet more diverse, the need for standardized representation formats becomes yet more evident. The evaluation framework described in this paper is a first step towards building an upper ontology of features and annotations.

The task of collecting and studying the datasets described in this paper has turned out to be more difficult than might have been anticipated due to problems with resource identification, retrieval and verification. The use of DOIs (document object identifier) to identify resources, and specialized sites which grant long-term storage and access (such as Zenodo) and machine-readable provenance and license information would partially solve these problems.

New initiatives such as the Spanish Retele network[42] address the same problems by gathering and organizing resources – and trying to make them easier to find. Additionally, established options like LDC, ELRA or MetaShare can be used to share resources more easily and efficiently. A normalized way to describe a corpus, such as the evaluation framework proposed here, would facilitate corpora filtering and selection.

Corpora used for sentiment analysis use annotations which are not adequately described; as a consequence, this hinders the unambiguous interpretation of categories and their comparison. We believe that the guidelines used for a corpus's annotation should be published along with each corpus, which would therefore facilitate automatic emotion mapping and enable corpora to be compared more effectively. Well-thought out methodologies for corpus creation in this sector could benefit from the potential of ontology-based models,

[42] `http://catalogo.retele.linkeddata.es`

interoperable metadata description and data repositories which offer long-term preservation.

References

1. Amigó, E., Carrillo de Albornoz, J., Chugur, I., Corujo, A., Gonzalo, J., Martín, T., Meij, E., de Rijke, M., Spina, D.: Overview of RepLab 2013: Evaluating Online Reputation Monitoring Systems. In: Proceedings of the Fourth International Conference of the CLEF initiative, pp. 333–352 (2013)
2. Atserias, J., Casas, B., Comelles, E., González, M., Padró, L., Padró, M.: FreeLing 1.3: Syntactic and semantic services in an open-source NLP library. In: Proceedings of LREC, vol. 6, pp. 48–55 (2006)
3. Boldrini, E., Balahur, A., Martínez-Barco, P., Montoyo, A.: Using EmotiBlog to annotate and analyse subjectivity in the new textual genres. Data Mining and Knowledge Discovery **25**(3), 603–634 (2012)
4. Breslin, J.G., Decker, S., et al.: SIOC: an approach to connect web-based communities. International Journal of Web Based Communities **2**(2), 133–142 (2006)
5. Brooke, J., Tofiloski, M., Taboada, M.: Cross-linguistic sentiment analysis: From english to spanish. In: Proceedings of the International Conference RANLP-2009, pp. 50–54. Association for Computational Linguistics, Borovets, Bulgaria (2009)
6. Cámara, E.M., Cumbreras, M.Á.G., Román, J.V., Morera, J.G.: Tass 2015–the evolution of the spanish opinion mining systems. Procesamiento del Lenguaje Natural **56**, 33–40 (2016)
7. Cambria, E., Livingstone, A., Hussain, A.: The Hourglass of Emotions, pp. 144–157. Springer Berlin Heidelberg, Berlin, Heidelberg (2012)
8. Cochrane, T.: Eight dimensions for the emotions. Social Science Information **48**(3), 379–420 (2009)
9. Cruz, F.L., Troyano, J.A., et al.: Clasificación de documentos basada en la opinión: experimentos con un corpus de crıticas de cine en espanol. Procesamiento de Lenguaje Natural **41**, 73–80 (2008)
10. Cunningham, H., et al.: Getting More Out of Biomedical Documents with GATE's Full Lifecycle Open Source Text Analytics. PLOS Computational Biology **9**(2), 1–16 (2013)
11. Ekman, P., Friesen, W.V., Ellsworth, P.: Emotion in the human face: Guidelines for research and an integration of findings. Pergamon Press (1972)
12. Fontaine, J.R.J., Scherer, K.R., Roesch, E.B., Ellsworth, P.C., Fontaine, J.R.J., Scherer, K.R., Roesch, E.B., Ellsworth, P.C.: The World of Emotions Is Not. Psychological Science **18**(12), 1050–1057 (2007)
13. Garcia-Moya, L., Anaya-Sanchez, H., Berlanga-Llavori, R.: Retrieving product features and opinions from customer reviews. IEEE Intelligent Systems **28**(3), 19–27 (2013)
14. Hepp, M.: Goodrelations: An ontology for describing products and services offers on the web. In: International Conference on Knowledge Engineering and Knowledge Management, pp. 329–346. Springer (2008)
15. Jiménez-Zafra, S.M., Martín-Valdivia, M.T., Maks, I., Izquierdo, R.: Analysis of patient satisfaction in dutch and spanish online reviews. Procesamiento del Lenguaje Natural **58**(0), 101–108 (2017)
16. Jiménez-Zafra, S.M., Martín-Valdivia, M.T., Molina-González, M.D., Ureña-López, L.A.: Corpus annotation for aspect based sentiment analysis in medical domain. In: Proceedings of the 2nd International Workshop on Extraction and Processing of Rich Semantics from Medical Texts (2017)
17. Jiménez-Zafra, S.M., Martín-Valdivia, M.T., Molina-González, M.D., Ureña-López, L.A.: Relevance of the SFU ReviewSP-NEG corpus annotated with the scope of negation for supervised polarity classification in Spanish. Information Processing & Management **54**(2), 240 – 251 (2018). DOI https://doi.org/10.1016/j.ipm.2017.11.007
18. Lövheim, H.: A new three-dimensional model for emotions and monoamine neurotransmitters. Medical Hypotheses **78**(2), 341–348 (2012)

19. Marcheggiani, D., Täckström, O., Esuli, A., Sebastiani, F.: Hierarchical multi-label conditional random fields for aspect-oriented opinion mining. In: Advances in Information Retrieval, pp. 273–285. Springer (2014)
20. Martí, M.A., Martín-Valdivia, M.T., Taulé, M., Jiménez-Zafra, S.M., Nofre, M., Marsó, L.: La negación en español: análisis y tipología de patrones de negación. Procesamiento del Lenguaje Natural **57**, 41–48 (2016)
21. Martín-Valdivia, M.T., Martínez-Cámara, E., Perea-Ortega, J.M., Ureña-López, L.A.: Sentiment polarity detection in spanish reviews combining supervised and unsupervised approaches. Expert Systems with Applications **40**(10), 3934–3942 (2013)
22. Martínez-Cámara, E., Martín-Valdivia, M.T., Ureña-López, L.A.: Opinion Classification Techniques Applied to a Spanish Corpus, pp. 169–176. Springer Berlin Heidelberg, Berlin, Heidelberg (2011). DOI 10.1007/978-3-642-22327-3_17
23. Martínez-Cámara, E., Martín-Valdivia, M.T., et al.: Polarity classification for Spanish tweets using the COST corpus. Journal of Information Science **41**(3), 263–272 (2015). DOI 10.1177/0165551514566564
24. Mehrabian, A.: Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in Temperament. Current Psychology **14**(4), 261–292 (1996)
25. Molina-González, M.D., Martínez-Cámara, E., et al.: Cross-domain sentiment analysis using Spanish opinionated words. In: Proceedings of NLDB, pp. 214–219 (2014). DOI 10.1007/978-3-319-07983-7_28
26. Nakamura, A.: Kanjo hyogen jiten. Tokyodo Publishing (1993)
27. Navas-Loro, M., Rodríguez-Doncel, V.: Oeg at tass 2017: Spanish sentiment analysis of tweets at document level (2017)
28. Navas-Loro, M., Rodríguez-Doncel, V., Santana-Pérez, I., Fernández-Izquierdo, A., Sánchez, A.: Mas: A corpus of tweets for marketing in spanish. In: A. Gangemi, A.L. Gentile, A.G. Nuzzolese, S. Rudolph, M. Maleshkova, H. Paulheim, J.Z. Pan, M. Alam (eds.) The Semantic Web: ESWC 2018 Satellite Events, pp. 363–375. Springer International Publishing, Cham (2018)
29. Navas-Loro, M., Rodríguez-Doncel, V., Santana-Perez, I., Sánchez, A.: Spanish Corpus for Sentiment Analysis towards Brands. In: Proc. of the 19th Int. Conf. on Speech and Computer (SPECOM), pp. 680–689 (2017)
30. Periñán-Pascual, C., Arcas-Túnez, F.: A knowledge-based approach to social sensors for environmentally-related problems. In: Intelligent Environments 2017: Workshop Proceedings of the 13th International Conference on Intelligent Environments, vol. 22, p. 49. IOS Press (2017)
31. Plaza-Del-Arco, F.M., Martín-Valdivia, M.T., et al.: COPOS: Corpus of patient opinions in Spanish. Application of sentiment analysis techniques. Procesamiento de Lenguaje Natural **57**, 83–90 (2016)
32. Plutchik, R.: The nature of emotions: Human emotions have deep evolutionary roots. American Scientist **89**(4), 344–350 (2001)
33. Pontiki, M., Galanis, D., Papageorgiou, H., Androutsopoulos, I., Manandhar, S., AL-Smadi, M., Al-Ayyoub, M., Zhao, Y., Qin, B., De Clercq, O., Hoste, V., Apidianaki, M., Tannier, X., Loukachevitch, N., Kotelnikov, E., Bel, N., Jiménez-Zafra, S.M., Eryiğit, G.: Semeval-2016 task 5: Aspect based sentiment analysis. In: Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016), pp. 19–30. Association for Computational Linguistics, San Diego, California (2016)
34. Rangel, F., Rosso, P., Reyes, A.: Emotions and Irony per Gender in Facebook. In: Proceedings of Workshop ES3LOD, LREC-2014,, pp. 1–6 (2014)
35. Reyes, A., Rosso, P.: Mining subjective knowledge from customer reviews: A specific case of irony detection. In: Proceedings of the 2nd workshop on computational approaches to subjectivity and sentiment analysis, pp. 118–124. Association for Computational Linguistics (2011)
36. Rivera Pastor, R., Tarín Quirós, C., Villar García, J.P., Badía Cardús, T., Melero Nogués, M.: Language equality in the digital age - Towards a Human Language Project (2017)
37. Roberto, J.A., Martí, M.A., Llorente, M.S.: Análisis de la riqueza léxica en el contexto de la clasificación de atributos demográficos latentes. Procesamiento del Lenguaje Natural **48**(0), 97–104 (2012)

38. Roberto, John A; Salamó, Maria; M. Antònia, M.: Clasificación automática del registro lingüístico en textos del español: un análisis contrastivo. LinguaMática **5**(1), 59–67 (2013)
39. Rodriguez-Doncel, V., Labropoulou, P.: Digital Representation of Rights for Language Resources. In: Proceedings of the 4th Workshop on Linked Data in Linguistics (LDL-2015), ACL-IJCNLP 2015, pp. 49–58 (2015)
40. Román, J.V., Morera, J.G., Cámara, E.M., Zafra, S.M.J.: Tass 2014 - the challenge of aspect-based sentiment analysis. Procesamiento del Lenguaje Natural **54**, 61–68 (2015)
41. Rosso, P., Rangel, F.: Author Profiling in Social Media: The Impact of Emotions on Discourse Analysis, pp. 3–18. Springer International Publishing, Cham (2017). DOI 10.1007/978-3-319-68456-7_1
42. Russell, J.: A circumplex model of affect. Journal of Personality and Social Psychology **39**(6), 1161–1178 (1980)
43. Sánchez Rada, J.F., Torres, M., et al.: A linked data approach to sentiment and emotion analysis of twitter in the financial domain. In: 2nd International Workshop on Finance and Economics on the Semantic Web (2014)
44. Shaver, P., Schwartz, J., et al.: Emotion knowledge: Further exploration of a prototype approach. Journal of personality and social psychology **52**(6), 1061–1086 (1987). DOI 10.1037/0022-3514.52.6.1061
45. Taboada, M., Brooke, J., Tofiloski, M., Voll, K., Stede, M.: Lexicon-based methods for sentiment analysis. Computational linguistics **37**(2), 267–307 (2011)
46. Tomkins, S.: Affect imagery consciousness: Volume I: The positive affects, vol. 1. Springer publishing company (1962)
47. Vazquez, K.L., Tovar, M., Vilariño, D., Beltrán, B.: Un algoritmo para detectar la polaridad de opiniones en los dominios de laptops y restaurantes. Advances in Intelligent Technologies and its Applications pp. 91–98 (2016)
48. Vilares, D.: Sentiment analysis for reviews and microtexts based on lexico-syntactic knowledge. In: FDIA'13, pp. 38–43 (2012)
49. Vilares, D., Alonso, M.A., Goméz-Rodríguez Carlos: A syntactic approach for opinion mining on Spanish reviews. Natural Language Engineering **1**(1), 1–26 (2013)
50. Villena-Román, J., García-Morera, J., Lana-Serrano, S., González-Cristóbal, J.C.: Tass 2013 - a second step in reputation analysis in spanish. Procesamiento del Lenguaje Natural **52**, 37–44 (2014)
51. Villena-Román, J., Lana-Serrano, S., Martínez-Cámara, E., González-Cristóbal, J.C.: Tass-workshop on sentiment analysis at sepln. Procesamiento del Lenguaje Natural **50**, 37–44 (2013)
52. Wiebe, J., Wilson, T., Cardie, C.: Annotating expressions of opinions and emotions in language. Language resources and evaluation **39**(2), 165–210 (2005)
53. Zafra, S.M.J., Berardi, G., Esuli, A., Marcheggiani, D., Martín-Valdivia, M.T., Fernández, A.M.: A multi-lingual annotated dataset for aspect-oriented opinion mining. In: EMNLP, pp. 2533–2538 (2015)