ISSN 2007-9737

TOM: Twitter Opinion Mining

Fernando M. Rodríguez, Sara E. Garza

Universidad Autónoma de Nuevo León (UANL),
Facultad de Ingeniería Mecánica y Eléctrica (FIME),
Mexico

fernando.aldape@gmail.com, sara.garzavl@uanl.edu.mx

Abstract. We present an opinion mining approach whose aim is to perform sentiment classification over microblogs in Spanish; since we use the Twitter microblog as a case study, this approach receives the name of Twitter Opinion Mining or TOM. To classify a comment as positive, negative, or neutral, TOM uses a term-counting strategy that sums the individual polarities of words and phrases contained in the comment. These polarities are obtained with an opinion lexicon that consists of weighted terms and valence shifters. Our lexicon not only includes generic terms translated from an English repository, but also more specific vocabulary from Twitter; this vocabulary is extracted by detecting adjectives and nouns from tweets with emoticons and trigrams that follow the "is-a" pattern. To assess TOM's quality, we measured precision, recall, and F_1 using a set of manually-classified tweets. Our results show high averages for each of these metrics, which were also used for comparing TOM against *Sentitext*, a tool for opinion mining in Spanish. The results for this comparison show that our approach outperforms this state of the art method.

Keywords. Opinion mining, sentiment analysis, lexicon, twitter, Spanish.

1 Introduction

Getting to know the opinion of customers about a product or service is valuable for companies and institutions. For example, for a construction company it would be helpful to know its potential client perception with regard to its costs, its quality, and offered house size; if a negative impression is detected, the company could make strategic decisions with respect to planning, human resources, and marketing, just to mention some.

With the intent of collecting opinions, companies usually invest in methods such as surveys, focus groups, and brand positioning analysis [10]. These studies take time and resources, both for being performed and for analyzing results, and this represents a disadvantage when only a limited budget is available. As a consequence, performing these studies, either directly or by means of outsourcing, is not always feasible [24].

Taking into account the aforementioned limitations, and at the same time knowing that the collection of consumers' perceptions is a vital task for the previously mentioned contexts, it is necessary to look for other alternatives. Social networks, for example, are spaces within the Internet where users can share ideas, expressions, and interests. These spaces are popular, provide real-time information, and have developed considerably over the past years. To illustrate this point, let us note that the social networking sites of Facebook are among Alexa's Top 10 Sites on a global scale[1], and that in countries such as the United States of America nearly 70% of Internet users have an account for social networks [14]. With respect to Mexico, Facebook has more than 85 million users and is the fifth country with more users for this social networking site[2].

In summary, social networks provide a world of information. From this world of information, it

[1] This information is from August 2019. Alexa's website is available at: `http://www.alexa.com/topsites`.

[2] "Mexico is the fifth country with more Facebook users in the world". Obtained from: `https://www.forbes.com.mx/mexico-el-quinto-pais-con-mas-usuarios-de-facebook-en-el-mundo/` (retrieved on August, 2019).

doi: 10.13053/CyS-23-4-2886

is possible to extract and analyze points of view, preferences, complaints, emotions, and thoughts [2, 30, 7], among other aspects, and this is cheaper than making a survey or gathering a group of people with a given profile [31]. All of this permits and motivates opinion mining over this kind of networks.

For the present work, we are particularly interested in opinion mining in Spanish and oriented towards microblogs. With regard to the language we have selected, it is well known that text analysis frequently needs to consider language-dependent aspects, and also that most available resources (lexicons, annotated data) are designed for the English language; consequently, one of our goals is to make a contribution to the Spanish language. To justify our choice for microblogs, let us note that this special kind of social network represents a hot research topic (specially Twitter, which is our case study) for its unique characteristics. One of these characteristics, probably the most important one, is that publications must be short (150-300 characters, typically). Even when this type of comment tends to be concise, which should facilitate opinion mining, it is also true that microblogs turn apart from the standard way of speaking, since they contain slang, neologisms, onomatopoeias, modisms, etc. This is one of the main challenges to address.

One of the most important microblogs nowadays is Twitter, which has almost 200 million users and where, on average, 60,000 comments get published daily. In this microblog, a user can publish short messages of at most 140 characters; these messages are called tweets; a user may, as well, follow other users to receive their tweets. Because of its popularity, availability, and sui generis features, we have selected Twitter as a case study.

Our approach for opinion mining in Spanish over microblogs is centered in the sentiment classification task, i.e., deciding whether a comment is positive, negative, or neutral. This approach, which we have named TOM for Twitter Opinion Mining, is based on the use of a lexicon and term-counting, i.e., balancing the negative and positive scores of the words in the text. We generate the lexicon semi-automatically by using comments with emoticons and trigrams with the "is a(n) [adjective/noun]" pattern. Currently, the lexicon has more than 68,000 sentiment words and phrases.

To evaluate TOM, we used a set of manually annotated tweets. We calculated precision, recall, and F-score, both globally and by sentiment class; results are also presented by activating and deactivating several features. We compared our results, as well, against Sentitext, which is another tool that performs opinion mining in Spanish. As we will see later, our results are favorable.

The rest of this document is organized as follows: Section 2 introduces notions that are necessary to understand our approach; Section 3 presents related work; Section 4 explains our approach; Section 5 describes experiments and results, and finally Section 6 presents conclusions and future work.

2 Background

Also known as sentiment analysis, opinion mining studies subjective expressions (reviews, comments, emotions, points of view, etc.) in media such as discussion forums, news, and blogs, just to mention some [13]. The starting point for this discipline consists of concretely defining what an opinion is. In our case, we shall use the definition by Go et al. [6]: "a personal positive or negative sensation".

Opinion mining has a set of related tasks, from which sentiment classification is the most common one. This task can be performed either at document level or at sentence level and consists of determining if an opinion expresses a positive, negative, or neutral sentiment (sometimes this last option is considered as equivalent to expressing no sentiment); this is also known as polarity detection, semantic orientation detection, or valence detection, and the neutral class is sometimes omitted.

Classification results are conventionally evaluated using precision, recall, and F-score. The aim of precision, on one hand, is to evaluate correctness and is calculated as:

$$p = \frac{TP}{TP + FP}, \qquad (1)$$

where TP is the number of elements correctly classified (true positives) and FP is the amount of elements that were labeled as belonging to the class, but did not actually belong (false positives).

The aim of recall, on the other hand, is to evaluate completeness and is calculated as:

$$r = \frac{TP}{TP + FN}, \qquad (2)$$

where FN is the number of elements that were not labeled as part of the class and did actually belong (false negatives).

The aim of F-score (F_1) is to integrate both precision and recall without privileging one over the other, as it is the harmonic mean of these two:

$$F_1 = 2 \times \left(\frac{p \times r}{p + r}\right). \qquad (3)$$

Sentiment classification, in general, can be *supervised* or *unsupervised*. Supervised approaches lean on machine learning methods (such as neural networks, support vector machines, Bayesian classifiers, etc.) and treat sentiment classification as a regular classification problem. As a result, these approaches require a *training set* (also known as *annotated data*).

Unsupervised sentiment classification, in contrast, comprises methods from different types (not necessarily disjoint), including the ones based in *lexicons*, the ones based in *term counting*, and the ones based in *natural language processing*. Term counting, particularly, consists of decomposing the document into words or phrases, detecting the polarity for each of these, and assigning the class with the dominant polarity. In the simplest case, if there are more positive than negative words, the document is positive and viceversa; if the amount of positive and negative words is equal, the document is neutral. Some works that implement this notion are the ones of Turney [26], Turney and Litmann [27], Dave et al. [5], and Kennedy and Inkman [9].

This last work also studies the impact of *contextual valence shifters*.

Valence shifters are terms that affect (decrease, increase, invert, or neutralize) the polarity of other terms [21]; for example, in the phrase “very pretty”, the word “very” is an intensifier. The two intrinsic aspects to a valence shifter are its *range of action* and its *intensification factor*. The former refers to the amount of words that will be affected (both before and after the valence shifter), while the latter refers to how much these words will be affected. Valence shifters can be included in lexicons via *rules*.

Another important task for opinion mining is the generation of opinion lexicons; this task consists of building resources that contain words, phrases, and rules that are *polar*, i.e., that express positive or negative sentiment. Lexicon generation can be manual, semi-automated, or fully automated. Manual generation is usually effective at the price of time and effort [12], while fully automated generation is less effective but has a wider coverage. Semi-automated generation combines the best of both worlds by automatically detecting polar words and then correcting mistakes manually. Both semi and fully automated lexicon generation usually rely on dictionaries and corpora.

Both supervised and unsupervised methods have pros and cons. For this reason, it is not uncommon to find *hybrid* methods in literature. A hybrid method, for example, could consist of using a lexicon as a starting point for creating a training set for a supervised method.

3 Related Work

The current section aims at describing outstanding works and approaches that are similar to ours (which presents a lexicon-based method for sentiment classification in Spanish over Twitter). In that sense, besides seminal work, we will be discussing approaches that analyze sentiment in Twitter, mine opinions in Spanish, and combine both aspects (note that the degree of similarity increases as the works are being explained). For each work, we describe the type of method (supervised, unsupervised), the language that the authors work with, and the repository being processed.

One of the most representative works is given by Turney [26], who classifies movie reviews by calculating the polarity of the phrases that compose these reviews and taking the average. If the average is positive, the review is classified as positive; otherwise, it is classified as negative. To calculate the polarity of a phrase, the PMI (pointwise mutual information) of the phrase and the word "excellent" is calculated, as well as the PMI between the phrase and the word "poor"; these two values are subtracted. Posterior works on this same approach employ latent semantic analysis (LSA) as an alternative to PMI and use a larger set of positive and negative words besides the two previously mentioned ones.

With respect to supervised approaches, the work by Pang et al. [19] classifies movie reviews utilizing three machine learning methods: a Bayesian classifier, a maximum entropy classifier, and support vector machines (SVM); unigrams, bigrams, adjectives, and part-of-speech (POS) tagging are used as features for the training phase. From the three classifiers, the support vector machines are the ones that obtain the best results (this technique, in general, has shown to be adequate for categorizing text).

Another seminal work is the one by Pak and Paroubek [18], which explores Twitter as a resource for opinion mining via the collection, analysis, and classification of tweets. The collection of positive and negative tweets is done through the use of emoticons, and the collection of neutral tweets is done with newspaper accounts.

With respect to analysis, it is centered in the linguistic and statistical aspects, and it establishes a comparison between subjective and objective text, as well as between positive and negative text; with these comparisons it was found, for example, that subjective text has more utterances and that verbs in past tense are more frequent in negative text, as they usually express regret or loss.

The third part of this work consists of classifying sentiment using the annotated tweets by Go et al. [6]; the best results were obtained by combining a Bayesian classifier with bigrams filtered via a salience metric proposed by the authors.

Following also the line of opinion mining in Twitter is the work by Zhang et al. [32], which combines the supervised and unsupervised approaches; this is achieved by using an opinion lexicon as starting point for training a classifier. Similarly, Go et al. [6] present an approach in which training sets are created assuming that emoticons indicate the sentiment class of the tweet (*distant supervision*); the best results are obtained with maximum entropy and taking both unigrams and bigrams as features. Furthermore, *query terms* are used to identify those tweets that are to be classified; the implemented protoype is available in `http://www.sentiment140.com` (requires a Twitter account). This prototype has recently been adapted for Spanish.

With respect to works focused on opinion mining in Spanish, one of the first approaches is given by Cruz et al. [4]; this work intends to reproduce the approach originally proposed by Turney by classifying movie reviews from the site `Muchocine`. The authors also propose a supervised method to define a *threshold* under which an opinion is negative and, otherwise, is positive.

The work by Brooke et al. [3] generates different lexicons for opinion mining in Spanish. Having previously created SO-CAL [25] for the English language, their first alternative consists of translating this lexicon with two distinct dictionaries (one of them being `Google Translate`) and making manual corrections, while their second alternative consists of manually constructing the lexicon from scratch with reviews in Spanish; a third alternative they consider is combining the two previous approaches. These alternatives are evaluated, along with an SVM, over corpora in Spanish and English; the authors conclude that translations have a cost and that, on the contrary, investing in generating specific resources for the target language is worthwhile.

Following also a crosslingual approach is the work by Pérez et al. [20], which creates lexicons for the Spanish language based on manual and automatic resources originally created for the English language. Their proposed method works at the *concept* level, that is, considering that several words belong to one same concept and one same word can belong to different concepts; for this reason, their alternatives involve the use of WordNet. The first of these alternatives consists of a series of successive alignments: from words of a manually annotated resource (OpinionFinder) to SentiWordNet *synsets*; from SentiWordNet synsets to English WordNet synsets; from English WordNet synsets to Spanish WordNet synsets. The second alternative consists of additionally collecting SentiWordNet concepts that do not appear in OpinionFinder. Both lexicons are evaluated using concept vectors (created with novels and the Spanish Wikipedia) to train an SVM. While the first lexicon achieves a greater precision, the second one achieves a greater coverage.

Other oustanding lexicon-based works include the one proposed by Kanayama and Nasukawa [8]. This work describes the automatic generation of lexicons for the Japanese language, and the method consists of detecting polar atoms, to which polarities are assigned according to context; polar atoms from an English lexicon are used as seeds, and these are translated automatically. The work is focused in photographic camera reviews.

In contrast with the previous works, which rely on the English language, the approach by Mellebeek et al. [15] considers non-expert annotations in Spanish for creating training sets. Their methodology includes the use of *Amazon's Mechanical Turk* to find users who are willing to classify (in exchange for a payment) the sentiment of a group of sentences in Spanish, which belong to the automobile reviews of the `ciao.es` site.

More similar to our approach is the work by Vilares et al. [29], which presents a hybrid technique for sentiment classification in Spanish over Twitter. The authors start with a lexical base that consists of several dictionaries in Spanish, which is used to calculate the polarity of nouns, verbs, adjectives, and adverbs. The treatment of negation and adversative subordinate clauses is then added to the approach by means of heuristics and dependency trees (an example of an adversative subordinate clause is “It's good, but I don't like it.”); this treatment is detailed in a previous work by the authors [28] developed over the SFU Spanish Review Corpus.

The previously mentioned polarity of terms is used as a feature to train a classifier, just as the number of positive and negative terms, the number of POS tags, the number of dependency types, and a *bag of words* with discriminant terms. The classifier was tested using the annotated corpus of tweets from the TASS (*Taller de Análisis del Sentimiento de la Sociedad Española para el Procesamiento de Lenguaje Natural*). With respect to the bag of words, it is also used to adapt the generic dictionaries to a specific domain, either by modifying the polarity of existing terms or by adding new ones.

Another highly related work is the one by Moreno et al. [17], which presents *Sentitext*. This tool performs sentiment classification in Spanish using a lexicon of almost 30,000 terms that are weighted in the range $[-2, 2]$; the lexicon, which was collected from OpenOffice, includes not only words but also phrases (currently 17,000) and rules regarding valence shifters. To calculate the polarity of an opinion, the weighted arithmetic mean is used to avoid that the document's length or the concentration of polar terms yields erroneous results.

Since the approach has initially been tested over hotel reviews (obtained from `Tripadvisor`), Sentitext assigns a number of *stars* to indicate how positive or negative the sentiment is; the range is 0-10 stars, where 0 indicates a very negative review, 5 a neutral one, and 10 a very positive one. The results over 100 reviews were calculated using a margin of error with respect to the actual number of stars; a 90% of accuracy was obtained.

The best results were obtained with positive reviews. In a posterior work [16], the polarity calculation was adapted for Twitter and the modifications were evaluated over the TASS corpus; in contrast with the results obtained with the hotel reviews, there is a considerable decrease in accuracy. The authors conclude that these results are due to the short length of tweets.

As we will see later, our approach can be distinguished in the following aspects: the form of collecting polar terms, the combined used of Twitter and generic lexicons, and the form of calculating polarity. Let us explain our approach more thoroughly.

4 Approach: Twitter Opinion Mining

Our aim consists of performing sentiment classification over a set of comments, i.e., detecting the polarity class (positive, negative, or neutral) for each one of these. Formally, given a set $\mathbb{C} = \{\mathcal{C}_1 \ldots \mathcal{C}_n\}$ of comments and a set $P = \{\texttt{POS}, \texttt{NEG}, \texttt{NEU}\}$ of polarity classes, we are interested in obtaining a function $f : \mathbb{C} \rightarrow P$, where for each pair $(\mathcal{C}_i, p) \in f$, the polarity class p is the correct one given the features of $\mathcal{C}_i$. As we will see in Section 5, we consider that p is the correct class for $\mathcal{C}_i$ when this pair is also found in a *reference set*, where this reference set can be a manual classification of the comments.

4.1 Lexicon Generation

Our lexicon contains weighted words and phrases, just as rules for valence shifting. The generation of this lexicon is done in three stages: (1) core construction, (2) expansion, and (3) weight assignment.

It is important to mention that this generation process is language and domain-independent; it could, therefore, be used in other contexts.

To construct the lexicon's core, which is our first stage, we collect a repository of comments (in our case, we also use Twitter). From these comments, we select those that contain emoticons, assuming that this indicates sentimental content [22]; the emoticons used can be seen in Table 1. From the obtained comments, we extract, via regular expressions, those *trigrams* (i.e., sequences of three words) that follow the pattern "is(are) a(n) [adjective/noun]"; the adjectives or nouns of these trigrams are considered *candidates* to be part of the lexicon, and are examined manually to leave only those that are actually polar terms. Let us note that the detection of this kind of trigrams allows to collect, among others, regional terms, slang, and modisms (e.g. "amors", "fregón", "nice").

Table 1. Emoticons used

Positive	:) :-) :D :-D
Negative	:(:-(:@ :-@

The core of our lexicon has 2,212 words: 1,065 of them are positive (48%) and 1,147 are negative (52%). To achieve a higher coverage, we expand this core by the inclusion of generic words, conjugations, phrases, and valence shifters, which is our second stage. With regard to the inclusion of generic words, we used the list of polar words by Liu [3] [11], which contains 6,800 words in English. These words are translated using Google Translate[4], which as we have seen previously, has already been used for crosslingual works. Those words whose sense is lost in translation are manually eliminated.

To conjugate verbs and obtain plural forms, we use the Freeling[5] tool [1], which has also been used in other works. This tool allows to carry out language analysis in Spanish and is free.

With respect to those phrases added to the lexicon, their purpose is to solve cases of ambiguity; for example, in Spanish the word "vale" in the phrase *me vale* ("I don't care") has a negative polarity, while in the phrase *vale la pena* ("it's worth it") it has a positive polarity. In that sense, our lexicon includes both exact phrases and *patterns*, where we call "pattern" to a dynamic phrase with variable words. An example of a pattern would be "*ni* ___, *ni dejan* ___ ", where the blank spaces admit a series of possible verbs: "*ni* trabajan, *ni dejan* trabajar" (you neither work, nor let others work), "*ni* hablan, *ni dejan* hablar" (you neither talk, nor let others talk).

With respect to valence shifters, we have included several common ones, such as *muy* (very), *mucho* (much), *poco* (little), *no*, and *realmente* (really). The range of action is of one word, either to the left, to the right, or both sides. If the shifter is an amplifier, the weight of the word is *doubled*, and if it is a diminisher, it is reduced by half. For example, if the polarity of a comment that contains the phrase *muy bonito* ("very beautiful") and the weight of "bonito" is 2.0, this weight will turn to 4.0 for the amplifier "muy".

In the specific case of negation, the range of action is of three subsequent words and the factor

[3] Available at: `http://www.cs.uic.edu/ liub/FBS/sentiment-analysis.html`

[4] Available at: `http://translate.google.com`

[5] Available at: `http://nlp.lsi.upc.edu/freeling/`

of intensification *cancels* the weight of these words if they are not very negative; otherwise, 1.0 is added to the current weight of the word. A formal description of the shifters can be seen in Table 2.

The third stage of our lexicon's generation process consists of word weight assignment, which is done manually (automatic assignment with a learning mechanism is left as future work). The assigned weights are in the range $[-3, 3]$, and each one of these weights is assigned according to the criteria shown in Table 3. Taking this into account, our lexicon, currently counts with 68,639 terms (words and phrases), from which 23,778 are positive (≈ 35%) and the rest are negative. Until now, the lexicon has 56 valence shifters.

4.2 Sentiment Classification

As we have mentioned previously, our algorithm for classifying the sentiment of comments is based on term-counting; as a result, we make a balance among the weights of the words that compose the comment and check whether this balance (which we will refer to as "polarity calculation") is mostly positive, negative, or neutral. Let us now explain with more detail this process.

The first step consists of pre-processing. This includes, on one hand, the use of delimiters (e.g. `,`, `;`, `?`) to eliminate questions, as these do not express an opinion. Furthermore, we divide the comment into words (i.e., create a bag of words) and carry out spelling correction.

For polarity calculation and posterior sentiment classification, we sum the weights of the words and phrases from comment $\mathcal{C}_i$ that have been found in the lexicon. We should consider three points: each phrase counts as a single word, the weight of each word can be affected by a valence shifter, and valence shifters have no weight (they only affect other words).

Formally, let $\mathcal{C}_i = F_i \cup T_i \cup M_i$, where F_i is the set of phrases in the comment, T_i is the set of words in the comment, and M_i is the set of valence shifters in the comment; note that these sets are disjoint, and that at least F_i or T_i must be non-empty. We first define a function $\omega(x)$ that obtains the corresponding weight of element $x \in F_i, T_i$ in the lexicon; if x is not found in the lexicon, $\omega(x) = 0$. Then, we define a function $\phi(m)$ that obtains the intensification factor for shifter $m \in M_i$ and, finally, a function $\mu(t)$ that obtains the valence shifter m that corresponds to word $t \in T_i$; if t is not altered by any shifter, $\phi(\mu(t)) = 1$. The polarity p of comment $\mathcal{C}_i$ is then given by:

$$p_i = \sum_{f \in F_i} \omega(f) + \sum_{t \in T_i} \phi(\mu(t)) \cdot \omega(t). \tag{4}$$

With respect to classification, the criterion to use (as we have seen with term-counting) is simple: if the sum of weights p_i is positive, then the comment is classified as positive; if it is negative, then the comment is classified as negative. However, this criterion is slightly modified to include the neutral class; in that sense, we make use of a range $[\alpha, \beta]$ inside of which a comment is neutral and outside of which it will be either positive (above β) or negative (below α) [9]. Formally, for a comment $\mathcal{C}_i$ with polarity p_i:

$$\text{class}(\mathcal{C}_i) = \begin{cases} \texttt{NEG} & \text{if } p_i < \alpha, \\ \texttt{NEU} & \text{if } \alpha \leq p_i \leq \beta, \\ \texttt{POS} & \text{if } p_i > \beta, \end{cases} \tag{5}$$

In our case, $\alpha = 0$ and $\beta = 1$.

5 Experiments and Results

To evaluate TOM, we calculated precision, recall, and F_1 over a set of comments in Spanish, which were extracted from Twitter. We present results both at a global scale and also by activating and deactivating several of TOM's features (e.g. the use of phrases or valence shifters). Moreover, with the intent of comparing our approach against the state of the art, we performed a comparison against Sentitext, which was previously described as one of the few tools that performs opinion mining in Spanish (Section 3). Let us describe data preparation, results, and the corresponding discussion.

Table 2. Valence Shifters. For the range of action, i represents the shifter position. For the operation, w is the current weight and w' is the new weight

	Range of action	**Operation**
Amplifier/ Diminisher	$(i-1)$ and/or $(i+1)$	$w' = 2w$ $w' = 0.5w$
Negation	$[i+1, i+3]$	$w' = \begin{cases} w+1 & \text{if } w < 1 \\ 0 & \text{otherwise} \end{cases}$

Table 3. Weight assignment criteria

Weight	**Description**	**Example**
$+3$	Positive sensation + compliment + praise	He is excellent in his work.
$+2$	Positive sensation + compliment	You are a good musician.
$+1$	Positive sensation	He obtained an award.
-1	Negative sensation	He had an accident.
-2	Negative sensation + insult	He was fired for being corrupt.
-3	Negative sensation + insult + humilliation	He has a mediocre attitude.

5.1 Dataset

The dataset for running experiments contains 3,000 tweets in Spanish. This dataset was obtained from a larger one, which contains approximately 40 million tweets and 19,000 users; this larger dataset, which has been used by the authors for other data mining tasks [23], was obtained by searching for tweets that contained one or more words related to “Monterrey” (a Northeastern city of Mexico, hometown of the authors) or its neighbor cities. The utilized words for searching are in Table 4.

From the resulting tweets, a *seed set* of 100 users was generated, and this set was expanded in a breadth-first fashion by obtaining the contacts from the seed users, and then the contacts from these contacts, and so on; with the intent of ensuring that the tweets were in Spanish, only those users whose profiles matched with the cities in Table 4 were considered. From each user, the 1,000 most recent tweets were collected (or all published tweets if an inferior quantity was available). From these tweets, a random sample of 6,000 was taken for TOM. With the purpose of having diversity in writing styles, each tweet belongs to a different user.

From the random sample of 6,000 tweets, half was used to create the lexicon (Section 4) and half was used for evaluating. These tweets were manually classified by three annotators whose mother language is Spanish; to avoid bias, each annotator classified independently by following his own criteria.

Available categories for classifying each tweet were: *Positive*, *Negative*, *Neutral*, *Sarcasm*, *Not in Spanish*, and *I don’t know*. Tweets falling into the last three categories were discarded, along with those tweets that were not unanimously classified by the annotators.

Taking this into account, the final sample for evaluation was composed of 1,147 tweets ($\approx$ 40% of the 3,000 initial tweets), from which 19% was classified as positive (220 tweets), 20% as negative (229), and 61% as neutral (698). Considering that the annotators only coincided 30% of the time, it seems interesting to note that sentiment classification in a corpus such as Twitter seems a complicated task not only for automated approaches, but even for humans.

Table 4. Words used for tweet seach

Apodaca	Escobedo	Guadalupe
Monterrey	Monterrey, N.L.	Mty
Santa Catarina	SanNicolas	San Nicolas
	San Pedro	

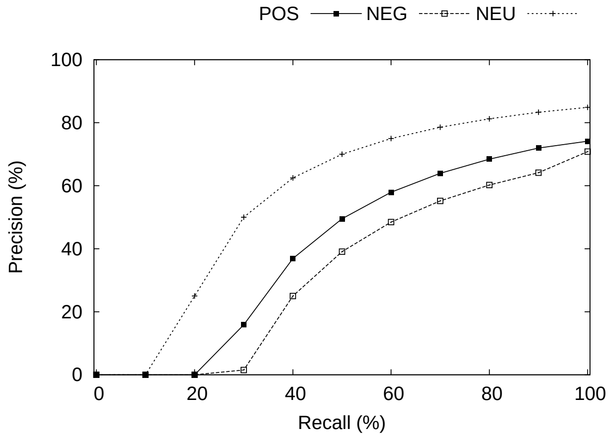


Fig. 1. Precision vs. recall

5.2 Results and Discussion

Table 5 presents the obtained results (in general and per sentiment class) by evaluating TOM with the previously described dataset. Figure 1 depicts, per class, precision with respect to different levels of recall. As we can see, the best results were obtained with neutral tweets, whereas the negative tweets represented the hardest case (this last finding agrees with the results obtained by Moreno et al. for Sentitext [17]). We believe this may be due to the use of sarcasm and irony.

Table 5. Sentiment classification results

	Precision (%)	**Recall** (%)	$\mathbf{F}_1$ (%)
Positive	75	78	76
Negative	71	69	70
Neutral	85	85	85
Global	**77**	**77**	**77**

With the intent of breaking down these results, as well as obtaining a perspective on the importance of TOM's individual components, we also calculated precision by progressively activating components. We started with a simple base and successively added components until getting the complete mechanism. Component addition proceeded as follows:

1. Liu's vocabulary,
2. Twitter vocabulary (instead of Liu's),
3. (1) + (2),
4. (3) + weight assignment,
5. (4) + spell checking,
6. (5) + valence shifters,
7. (6) + use of phrases,
8. (7) + question elimination.

Results are summarized in Table 6, which includes the precision obtained per component and the percentage that it improved (or worsened) when the component was added. As we can see, the most significant improvements were given by using the vocabulary extracted from Twitter ("is-a" trigrams), by assigning weights, and by combining Lui's lexicon with ours (Twitter). Moreover, let us note (although the improvement by itself was substantial) that most of the improvements happened with positive comments; on the contrary, precision went slightly down with neutral comments.

We consider that the use of the Twitter vocabulary impacts results, as it is more attached to the domain, since it includes words that go along with the writing style that is used in there. In comparison with this vocabulary, the one of Liu is more generic and, as seen already in other works, translations are not always efficient. However, it does not come as a surprise that the combination of both vocabularies into a single lexicon obtains a better precision, since both the generic and the domain-oriented aspects are being covered (we consider in this case, that the domain not only includes topics but also writing styles).

On the other hand, it seems logical to expect for weight assignment to contribute for a considerable improvement in precision, as this component increases the granularity of the lexicon and permits to generate a higher degree of separation between classes (let us think, for example, in the difference between the intensity of a bad word and a negative expression). With respect to the improvement per class, we can see that the improvements are mainly given in the positive and negative classes; this could be due to the fact that all components are oriented towards detecting tweets with a high polarity (either positive or negative).

Table 6. Precision (%) activating components (Δ = difference)

	Liu	**Twitter**	**Liu + Twitter**	**Weights**	**Spelling**	**V. Shifters**	**Phrases**	**Questions**	**Δ**
Positive	26	46	53	73	74	76	78	78	
		(+20)	*(+7)*	*(+20)*	*(+1)*	*(+2)*	*(+2)*	*(0)*	*+52*
Negative	50	59	59	68	69	70	70	70	
		(+9)	*(0)*	*(+9)*	*(+1)*	*(+1)*	*(0)*	*(0)*	*+20*
Neutral	87	89	86	84	84	84	83	85	
		(+2)	*(-3)*	*(-2)*	*(0)*	*(0)*	*(-1)*	*(+2)*	*-2*
Global	54	65	66	75	76	76	77	77	
		(+11)	*(+1)*	*(+9)*	*(+1)*	*(0)*	*(+1)*	*(0)*	*+23*

Table 7. Comparative results (T=TOM, S=Sentitext)

	Precision (%)		
	T	S	Δ
POS	74.5	56.5	***+18.1***
NEG	69.9	70.4	*-0.5*
NEU	85.4	84.0	*+1.4*
Global	80.3	74.7	*+5.6*

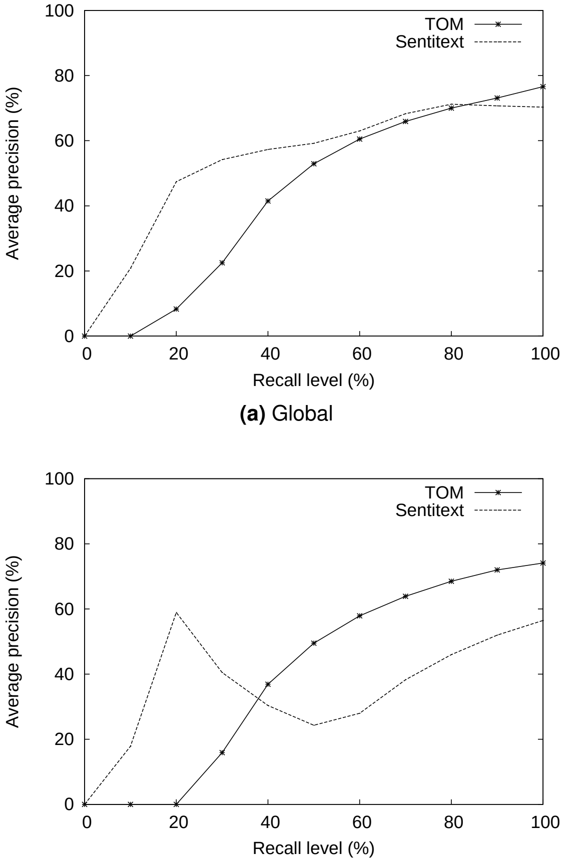
	Recall (%)		
	T	S	Δ
POS	77.4	75.5	*+1.9*
NEG	68.3	62.4	*+5.9*
NEU	85.0	78.5	***+6.6***
Global	80.3	74.7	*+5.6*

	$\mathbf{F}_1$ (%)		
	T	S	Δ
POS	75.9	64.6	***+11.3***
NEG	69.1	66.2	*+2.9*
NEU	85.2	81.1	*+4.1*
Global	76.7	70.6	*+6.1*

5.3 Comparison with Sentitext

The last part of our evaluation consists of showing that TOM is competitive with respect to the state of the art. Consequently, we compared TOM against a representative approach for opinion mining in Spanish: Sentitext.

This tool has several years of continuous development, is similar to TOM, and is one of the few that exists for the Spanish language.



(a) Global

(b) Positive

Fig. 2. Comparative results for the global and positive cases

The design and results for these experiments were performed with the original version used for hotel reviews [17], since this version was the one available by that time. Since Sentitext uses stars (0-10), where 0 indicates that the text is very negative and 10 that it is very positive, it was necessary to cast these results into the three sentiment classes (positive, negative, and neutral).

To have a uniform criterion of conversion, we considered that a score equal to 5 corresponded to neutral, a score greater than 5 corresponded to positive, and a score less than 5 corresponded to negative.

With respect to the dataset, we used the same dataset of the previous experiments (1,147 tweets); however, 31 tweets had to be discarded, since they produced an error when they were introduced to the Sentitext prototype.

Comparison results are shown in Table 7 and Figure 2. While globally both approaches had a similar performance, with TOM being better by a small margin, one of the most notable results is given by the positive class (Figure 2b), where TOM obtained a precision of 74.5% and Sentitext a precision of 56.5% (75.9% versus 65.6% with respect to F_1).

While it is true that TOM is oriented towards Twitter and Sentitext is originally intended for longer texts, it is also true that Sentitext has been refining its lexicon for years.

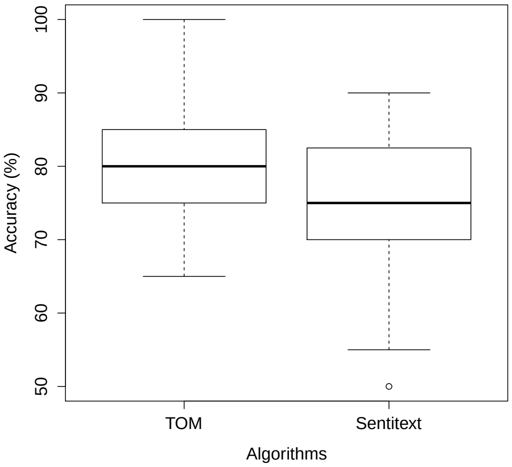


Fig. 3. Comparison with Sentitext

6 Conclusions and Future Work

We have presented an approach for opinion mining in Spanish, and we have used Twitter as a case study. This approach tackles, specifically, document sentiment classification (in our case, we consider a comment or tweet to be a document) with three classes: positive, negative, and neutral. The classification algorithm is based in the sum of the individual polarities of words and phrases contained in the comment. To obtain these polarities, we generate an opinion lexicon that contains weighted words, phrases, and valence shifters. An important part of the lexicon was extracted from Twitter by means of the detection of comments with emoticons and trigrams following the "is-a" pattern.

With respect to validation, we have evaluated TOM using a set of tweets. Precision, recall, and F_1 are more than satisfactory; when comparing against the Sentitext tool, TOM also shows competitive or better results (mainly with the positive class).

With regard to future work, it can be grouped under distinct aspects. One of them concerns improving the lexicon, which includes its expansion, refinement, and automation; furthermore, techniques such as deep learning can be incorporated. Another aspect consists of evaluating TOM in other contexts, such as review repositories. A third aspect consists in the use of the extracted tweets to generate resources in Spanish, such as the PMI for pairs of words.

References

1. **Atserias, J., Casas, B., Comelles, E., González, M., Padró, L., & Padró, M. (2006).** FreeLing 1.3: Syntactic and semantic services in an open-source NLP library. *Proceedings of Language Resources and Evaluation Conference*, volume 6, pp. 48–55.

2. **Balduzzi, M., Platzer, C., Holz, T., Kirda, E., Balzarotti, D., & Kruegel, C. (2011).** Abusing social networks for automated user profiling. *Recent Advances in Intrusion Detection*, Springer, pp. 422–441. `https://link.springer.com/chapter/10.1007/978-3-642-15512-3_22`.

3. **Brooke, J., Tofiloski, M., & Taboada, M. (2009).** Cross-Linguistic Sentiment Analysis: From English to Spanish. *Recent Advances in Natural Language Processing (RANLP) 2009*, pp. 50–54. https://www.aclweb.org/anthology/R09-1010.

4. **Cruz, F. L., Troyano, J. A., Enriquez, F., & Ortega, J. (2008).** Clasificación de documentos basada en la opinión: experimentos con un corpus de crıticas de cine en español. *Procesamiento de Lenguaje Natural*, Vol. 41, pp. 73–80. http://rua.ua.es/dspace/handle/10045/8067.

5. **Dave, K., Lawrence, S., & Pennock, D. M. (2003).** Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. *Proceedings of the 12th International Conference on World Wide Web*, ACM, pp. 519–528. http://doi.acm.org/10.1145/775152.775226.

6. **Go, A., Huang, L., & Bhayani, R. (2009).** Twitter sentiment classification using distant supervision. Technical Report CS224N, Stanford University.

7. **Goyal, A., Bonchi, F., & Lakshmanan, L. (2010).** Learning influence probabilities in social networks. *Proceedings of the 3rd ACM International Conference on Web Search and Data Mining*, ACM, pp. 241–250. http://doi.acm.org/10.1145/1718487.1718518.

8. **Kanayama, H. & Nasukawa, T. (2006).** Fully automatic lexicon expansion for domain-oriented sentiment analysis. *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, pp. 355–363. http://dl.acm.org/citation.cfm?id=1610075.1610125.

9. **Kennedy, A. & Inkpen, D. (2006).** Sentiment classification of movie reviews using contextual valence shifters. *Computational Intelligence*, Vol. 22, No. 2, pp. 110–125. https://doi.org/10.1111/j.1467-8640.2006.00277.x.

10. **Kotler, P., Kartajaya, H., & Setiawan, I. (2011).** *Marketing 3.0: From Products to Customers to the Human Spirit*. Wiley, New Jersey, USA.

11. **Liu, B. (2010).** *Handbook of Natural Language Processing*, chapter Sentiment Analysis and Subjectivity. Chapman & Hall, 2nd edition, pp. 627–666.

12. **Liu, B. (2012).** Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*, Vol. 5, No. 1, pp. 1–167. https://doi.org/10.2200/S00416ED1V01Y201204HLT016.

13. **Liu, B. & Zhang, L. (2012).** A survey of opinion mining and sentiment analysis. In **Aggarwal, C. C. & Zhai, C.**, editors, *Mining Text Data*. Springer US, pp. 415–463. https://doi.org/10.1007/978-1-4614-3223-4_13.

14. **Maeve Duggan, J. B. (2012).** The demographics of social media users - 2012. Descargado el 23 de Febrero de 2013.

15. **Mellebeek, B., Benavent, F., Grivolla, J., Codina, J., Costa-jussà, M. R., & Banchs, R. (2010).** Opinion mining of spanish customer comments with non-expert annotations on mechanical turk. *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, Association for Computational Linguistics, Stroudsburg, Pennsylvania, USA, pp. 114–121. http://dl.acm.org/citation.cfm?id=1866696.1866714.

16. **Moreno-Ortiz, A. & Hernández, C. P. (2012).** Lexicon-based sentiment analysis of Twitter messages in Spanish. *Procesamiento del lenguaje natural*, Vol. 50, pp. 93–100. http://rua.ua.es/dspace/handle/10045/27869.

17. **Moreno-Ortiz, A., Pineda-Castillo, F., & Hidalgo-García, R. (2010).** Análisis de Valoraciones de Usuario de Hoteles con Sentitext: un sistema de análisis de sentimiento independiente del dominio. *Procesamiento del lenguaje natural*, Vol. 45, pp. 31–39. http://rua.ua.es/dspace/handle/10045/14724.

18. **Pak, A. & Paroubek, P. (2010).** Twitter as a Corpus for Sentiment Analysis and Opinion Mining. *Language Resources and Evaluation*, European Language Resources Association, Valletta, Malta, pp. 1320–1326. https://lexitron.nectec.or.th/public/LREC-2010_Malta/pdf/385_Paper.pdf.

19. **Pang, B., Lee, L., & Vaithyanathan, S. (2002).** Thumbs up?: Sentiment classification using machine learning techniques. *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, ACM, New York, USA, pp. 79–86. https://doi.org/10.3115/1118693.1118704.

20. **Perez-Rosas, V., Banea, C., & Mihalcea, R. (2012).** Learning Sentiment Lexicons in Spanish. *LREC*, pp. 3077–3081. http://lrec.elra.info/proceedings/lrec2012/pdf/1081_Paper.pdf.

21. **Polanyi, L. & Zaenen, A. (2006).** Contextual valence shifters. In **Shanahan, J., Qu, Y., & Wiebe, J.**, editors, *Computing Attitude and Affect*

in Text: Theory and Applications, volume 20 of *The Information Retrieval Series*. Springer Netherlands, pp. 1–10. http://dx.doi.org/10.1007/1-4020-4102-0_1.

22. **Read, J. (2005).** Using emoticons to reduce dependency in machine learning techniques for sentiment classification. *Proceedings of the ACL Student Research Workshop*, Association for Computational Linguistics, pp. 43–48. http://dl.acm.org/citation.cfm?id=1628960.1628969.

23. **Rodríguez, F. M. (2013).** *Cuantificación del interés de un usuario en un tema mediante minería de texto y análisis de sentimiento*. Master's thesis, Universidad Autónoma de Nuevo León. http://eprints.uanl.mx/3679/.

24. **Sharma, G. & Singh, S. (2011).** Economic Analysis of Post-harvest Losses in Marketing of Vegetables in Uttarakhand. *Agricultural Economics Research Review*, Vol. 24, No. 2, pp. 309–315. https://ageconsearch.umn.edu/record/119384/.

25. **Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011).** Lexicon-based methods for sentiment analysis. *Computational linguistics*, Vol. 37, No. 2, pp. 267–307. https://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00049.

26. **Turney, P. (2002).** Thumbs up or thumbs down?: Semantic Orientation Applied to Unsupervised Classification of Reviews. *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, Association for Computational Linguistics, pp. 417–424. https://doi.org/10.3115/1073083.1073153.

27. **Turney, P. D. & Littman, M. (2003).** Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems (TOIS)*, Vol. 21, No. 4, pp. 315–346. http://doi.acm.org/10.1145/944012.944013.

28. **Vilares, D., Alonso, M. A., & Gómez-Rodríguez, C. (2013).** Clasificación de polaridad en textos con opiniones en español mediante análisis sintáctico de dependencias. *Procesamiento del lenguaje natural*, Vol. 50, pp. 13–20. http://rua.ua.es/dspace/handle/10045/27859.

29. **Vilares, D., Alonso, M. A., & Gómez-Rodríguez, C. (2013).** Una aproximación supervisada para la minería de opiniones sobre tuits en español en base a conocimiento lingüístico. *Procesamiento del Lenguaje Natural*, Vol. 51, pp. 127–134. http://rua.ua.es/dspace/handle/10045/30627.

30. **Wang, Y., Cong, G., Song, G., & Xie, K. (2010).** Community-based greedy algorithm for mining top-k influential nodes in mobile social networks. *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, pp. 1039–1048. http://doi.acm.org/10.1145/1835804.1835935.

31. **Zavišić, S. & Zavišić, Ž. (2011).** Social network marketing. *22nd CROMAR Congress*, pp. 1008–1019.

32. **Zhang, L., Ghosh, R., Dekhil, M., Hsu, M., & Liu, B. (2011).** Combining lexicon-based and learning-based methods for Twitter sentiment analysis. Technical Report 89, HP Laboratories. http://www.hpl.hp.com/techreports/2011/HPL-2011-89.pdf.

Article received on 26/01/2018; accepted on 20/08/2019. Corresponding author is Fernando M. Rodríguez.