Sentiment Analysis of Spanish Tweets Using a Ranking Algorithm and Skipgrams*

Análisis de sentimientos sobre tweets en castellano utilizando un algoritmo de ranking y skipgrams

Javi Fernández, Yoan Gutiérrez, José M. Gómez, Patricio Martínez-Barco, Andrés Montoyo, Rafael Muñoz
Departamento de Lenguajes y Sistemas Informáticos, Universidad de Alicante
{javifm,ygutierrez,jmgomez,patricio,montoyo,rafael}@dlsi.ua.es

Resumen: En este artículo presentamos nuestra contribución a la *Tarea 1* (clasificación de polaridad en 6 niveles) de la competición *TASS 2013*. Esta contribución está formada por dos aproximaciones diferentes: una versión modificada de un algoritmo de ranking (RA-SR) utilizando bigramas, y una nueva propuesta que utiliza un puntuador de skipgrams. Estas aproximaciones crean diccionarios de sentimientos capaces de mantener el contexto de los términos. Todas nuestras aproximaciones aparecen en los primeros 10 mejores resultados entre los sistemas presentados a la competición, y la combinación de ambos consigue llegar a la primera posición.
Palabras clave: análisis de sentimientos, minería de opiniones, generación de lexicones, aprendizaje automático, twitter, algoritmo de ranking, skipgrams

Abstract: In this paper, we present our contribution for the *Task 1* (6 levels polarity classification) of the *TASS 2013* competition. This contribution consists on two different approaches: a modified version of a ranking algorithm (RA-SR) using bigrams, and new proposal using a skipgrams scorer. These approaches create sentiment lexicons able to retain the context of the terms. All our approaches appear in the top 10 best results of the systems presented to the competition, and the combination of them reaches the first position.
Keywords: sentiment analysis, opinion mining, lexicon generation, machine learning, twitter, ranking algorithm, skipgrams

1 Introduction

Textual information has become one of the most important sources of data to extract useful and heterogeneous knowledge from. Texts can provide *factual information*, such as descriptions, lists of features, or even instructions, and *opinion-based information*, which would include reviews, emotions, or feelings. This subjective information can be expressed through different textual genres, such as blogs, forums, and reviews, but also through social networks and microblogs.

Twitter is a microblogging social network that has gained much popularity last years. This service enables its users to send and read text-based messages of up to 140 characters, known as *tweets*. This site can be a vast source of subjective information in real time; millions of users share opinions on different aspects of their everyday life. Extracting this subjective information has a great value for both general and expert users. For example, users can find opinions about a product they are interested in, and companies and public figures can monitor their online reputation. Traditional *Sentiment Analysis* (SA) can deal with this task; however, it is difficult to exploit it accordingly, mainly because of the short length of the tweets, the informality, and the lack of context. SA systems must be adapted to this face the challenges of this new textual genre.

* We would like to express our gratitude for the financial support given by the Department of Software and Computer Systems at the University of Alicante, the Spanish Ministry of Economy and Competitivity (Spanish Government) by the project grants TEXT-MESS 2.0 (TIN2009-13391-C04-01), LEGOLANG (TIN2012-31224), ATTOS (TIN2012-38536-C03-03), SAM (FP7-611312), and the Valencian Government (grant no. PROMETEO/2009/119).

Some international competitions related to the assessment of SA systems in Twitter have taken place. Some of those include *CLEF RepLab*[1] (Amigó et al., 2012), *Semeval 2013 Task 2*[2] (Kozareva et al., 2013), and *SEPLN TASS*[3] (Villena Román et al., 2013).

In this paper, we present our contribution for the *Task 1* of the *TASS 2013* competition[4] (Villena Román et al., 2013). This task consists on performing an automatic sentiment analysis to determine the global polarity of a set of tweets. The polarity is divided in 6 levels: *positive* (`P`), *strong positive* (`P+`), *negative* (`N`), *strong negative* (`N+`), *neutral* (`NEU`) and *no sentiment* (`NONE`). This task provides a training corpus with 7,219 tweets and a test corpus with 60,798 tweets. The distribution of polarities in these datasets is shown in Table 1. Participants must classify each message in the test set, and may use the training set to train and validate their models.

Polarity	**Train**	**Test**
`P`	1,232	1,488
`P+`	1,652	20,745
`N`	1,335	11,287
`N+`	847	4,557
`NEU`	1,483	1,305
`NONE`	670	21,416
Total	7,219	60,798

Table 1: TASS Dataset Distribution (in number of tweets)

Our contribution consist on two different approaches: a modified version of the *UMCC-DLSI-(SA)* system, used on the *Task 2* of the *Semeval 2013* competition; and new proposal using a skipgrams scorer. Both approaches generate a sentiment resource, and employ machine learning techniques to detect the polarity of a text. They are presented in detail in Sections 3 and 4 respectively. Subsequently, in Section 5 we show the assessment of our model in the competition. Finally, the conclusions and future work are presented in Section 6. The following Section 2 shows some relevant background related to this work.

[1]http://www.limosine-project.eu/events/
[2]http://www.cs.york.ac.uk/semeval-2013/task2/
[3]http://www.daedalus.es/TASS/
[4]http://www.daedalus.es/TASS2013/

2 Related Work

The use of sentiment resources has proven to be a necessary step for training and evaluating systems that implement sentiment analysis, which also include fine-grained opinion mining (Balahur et al., 2010). In order to build sentiment resources, several studies have been conducted. One of the first is the relevant work by (Hu and Liu, 2004) using lexicon expansion techniques by adding synonymy and antonym relations provided by WordNet (Miller and Fellbaum, 1998; Miller, 1993). Another one is the research described by (Liu, Hu, and Cheng, 2005; Hu and Liu, 2004) which obtained an Opinion Lexicon compounded by a list of positive and negative opinion words or sentiment words for English (around 6,800 words) and Spanish (around 1,500 words). A similar approach has been used for building WordNet-Affect (Strapparava and Valitutti, 2004) which expands six basic categories of emotion; thus, increasing the lexicon paths in WordNet.

Another well presented lexicon can be found it in (Pérez-Rosas, Banea, and Mihalcea, 2012), where 2,496 words in Spanish are annotated into two different lexicons. The first one is named *Full Strength Lexicon*, which is more robust, as it leverages manual sentiment annotations from the OpinionFinder lexicon (Wiebe, Wilson, and Cardie, 2005). The second one is called *Medium Strength Lexicon*, which leverages automatic annotations induced based on SentiWordNet (Esuli and Sebastiani, 2006).

Nowadays, many sentiment and opinion messages are provided by Social Media. In it, new expression manners characterise the communication streaming across the Social Medias. That reason is very important to us, because allow us retrieving available information in these medias for build sentiment resources of new type.

3 Classifier I: Ranking Algorithm with Unigrams and Bigrams

The first of our approaches creates a sentiment resource by adding lexical patterns, to generate a classifier that can deal with the challenge posted in the competition. In order to build this sentiment resource we use a method named *RA-SR* (using Ranking Algorithms to build Sentiment Resources) (Gutiérrez et al., 2013a) which is able to produce sentiment inventories based on senti-

semantic evidence, obtained after exploring text with annotated sentiment polarity information. Through this process, a graph-based algorithm is used to obtain auto-balanced values that characterise sentiment polarities. This method consists of three key stages: *i)* building contextual word graphs; *ii)* applying a ranking algorithm; and *iii)* adjusting the sentiment polarity values. These stages are shown in Figure 1.

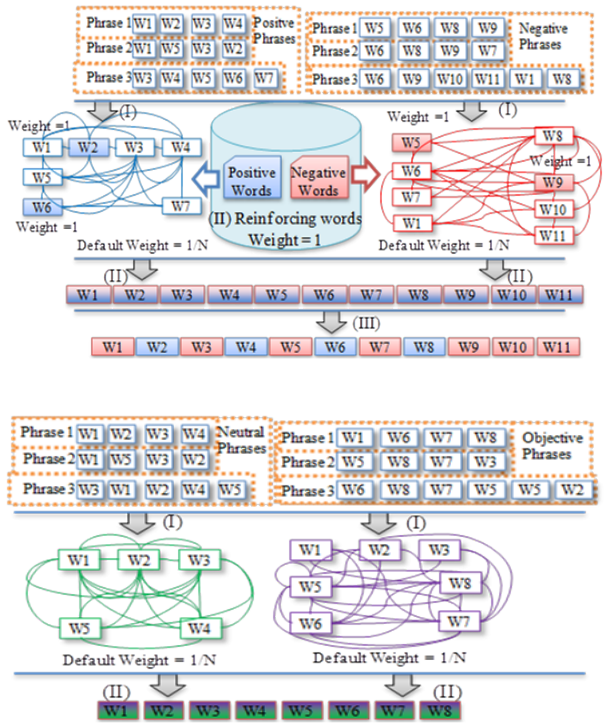

Figure 1: Resource walkthrough development process (RA-SR)

The main difference with the basic proposal described in (Gutiérrez et al., 2013a), which builds the contextual graphs representing words as the graph nodes, is that in this approach we consider not only words but also word *bigrams* to represent these graph nodes.

3.1 Sentiment Resource

The development of the sentiment resource starts by giving four corpora of annotated sentences: the first with neutral sentences, the second with objective sentences, the third with positive sentences, and the last with negative sentences.

Afterwards, a filtering process is applied to each sentence, making the following replacements using regular expressions: *i)* Internet addresses are replaced by the label `URL`, *ii)* emails are replaced by the label `MAIL`, and *iii)* entities and nicknames are replaced by the label `ENTITY`.

Subsequently, texts are preprocessed and tokenised using *Freeling 2.2*[5] (Padró et al., 2012), to obtain lemmatised unigrams and bigrams. It is important to remark that all lemmatised words are considered to select the elements that involve the nodes, without considering the part of speech as a filter. An example is shown in Figure 2.

Tweet:
"Conozco a alguien q es adicto al drama! Ja ja ja te suena d algo!"
↓
Unigrams:
conocer, a, alguien, q, ser, adicto, al, drama, !, ja, ja, ja, te, sonar, d, algo, !
↓
Bigrams:
conocer a, a alguien, alguien q, q ser, ser adicto, adicto al, al drama, drama !, ! ja, ja ja, ja ja, ja te, te sonar, sonar d, d algo, algo !

Figure 2: Example of unigrams and bigrams lemmatisation

Therefore, unigrams u and bigrams b conform the vertexes V of the RA-SR contextual graph G. This graph is represented as $G = (V, E)$, where $V = \{v_1, v_2, ..., v_n\}$ and $E = \{e_{1,1}, e_{1,2}, ..., e_{n,n}\}$ being $e_{i,j}$ an edge between v_i and v_j and n the number of unigrams and bigrams. The links between two edges are created associating all grams involved in each analysed text, so $e_{i,j} \in E$ and $v_i \in V$.

At this point, each unigram/bigram obtained in the sentiment resource has an associated value of *positivity*, *negativity* and *objectivity* obtained by applying the *PageRank* algorithm over lexical graphs that represent each polarity respectively. Note that RA-SR uses a lexicon repository of positive and negative well known words (see Table 4) to reinforce the contextual graphs. These three features are obtained by means of several normalisation equations that permit balancing the obtained values for each polarity perspective. On the other hand, objective and neutral scores are also obtained without any balancing. These five features we name *pos* (computed positivity), *neg* (computed

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

negativity), *obj_measured* (computed objectivity), *neu* (neutrality) and *real_obj* (objectivity). Note that, two objective scores are obtained by different ways, one as result of $1 - |negativity - positivity|$ and the other by applying PageRank over an objective contextual graph. A detailed explanation at word level can be found it in (Gutiérrez et al., 2013a) and (Gutiérrez et al., 2013b). Finally, the resource is formed by unigrams and bigrams, with five scores of polarities associated: *pos*, *neg*, *obj_measured*, *neu*, and *real_obj*.

3.2 Features

Features are obtained from the previously created resource. Each text is preprocessed using Freeling, as above-mentioned, obtaining its lemmatised unigrams and bigrams. Then, we sum all the corresponding values for each coincident gram between the analysed text and the sentiment resource, obtaining a single value of *pos*, *neg*, *obj_measured*, *neu* and *real_obj* for each text respectively. Finally, these values are normalised by dividing them by the number of grams in the text.

Other computed features are *pos_count*, *neg_count*, *obj_measured_count*, *obj_real_count* and *neu_count*. These features count each involved iteration for each feature type (*pos*, *neg*, *obj_measured*, *neu* and *real_obj*) respectively, where the value is greater than zero.

Other features have been obtained based on characteristics of the phrases, which can help with the definition on extreme cases. Feature *emotPos* corresponds to the number of positive emoticons, feature *emotNeg* is the number of negative emoticons, feature *exc* counts the number of exclamation marks, and feature *itr*, the number of interrogation marks. Features *cnp* and *cnn* represent the total of positive and negative words respectively implicated on each analysed text, identified using the sentiment lexicons described in Table 4. Finally, we introduce the feature *words_count* as a normalisation parameter, which acts as scalar regulator of the rest of the features, indicating the length of the analysed phrase (including bigrams). Table 2 summarises all these features, which will be used to create the polarity classifier employing a machine learning algorithm.

Feature	Description
pos *neg* *obj_measured* *real_obj* *neu*	Sums the respective value of each unigram/bigram
pos_count *neg_count* *obj_measured_count* *real_obj_count* *neu_count*	Counts the unigrams/bigrams where its respective value is greater than zero
cnp *cnn*	Counts the unigrams/bigrams contained in the Sentiment Lexicons for their respective polarities
emotPos *emotNeg*	Counts the respective emoticons
exc *itr*	Counts the respective marks
words_count	Counts the grams in the sentence

Table 2: Summary of features employed

3.3 TASS Implementation

To create the sentiment resource we used three different datasets, whose distribution is shown in Table 3:

- *TASS Train* (TASS). Represents all train tweets provided by TASS 2013 competition. Note that sentences annotated as `NONE` are considered as objective.
- *RAE Definitions* (RAE). Contains definitions from *Real Academia Española* for each word of two different resources, *Sentiment Lexicons in Spanish* (SLS) (Pérez-Rosas, Banea, and Mihalcea, 2012) and *MQPA Spanish* (Liu, Hu, and Cheng, 2005; Hu and Liu, 2004), presented on Table 4. It is important to remark that SLS can have two polarity classification for each word (`P` or `N`). Words where these classifications are different are ignored.
- *Twitter Emoticon Queries* (TEQ). This resource has been created specifically for this work. It contains positive and neg-

ative tweets obtained by searching the emoticons `:)` and `:(` on Twitter. Once we have retrieved the tweets from those queries, the ones containing the emoticon `:)` are considered positive, and the ones containing the emoticon `:(` are considered negative. Tweets containing both emoticons are ignored.

	TASS	**RAE**	**TEQ**	**Total**
P	1,232	7,827	49,289	60,000
P+	1,652			
N	1,335	9,507	48,311	60,000
N+	847			
NEU	1,483	-	-	1,483
NONE	670	-	-	670
Total	7,219	17,334	97,600	

Table 3: Corpora distribution (in number of texts)

	SLS	**MPQA**	**Total**
P	472	753	1,225
N	866	768	1,634
Total	1,338	1,521	2,859

Table 4: Sentiment lexicons distribution (in number of words)

Employing the annotated sentences provided by these corpora, we built the sentiment resource. Then, using the features described, a classifier is created using the Weka[6] (Hall et al., 2009) default implementation of the *Support Vector Machines* (SVM) algorithm. We chose this classifier due to its good performance in text categorisation (Sebastiani, 2002) and in previous works in sentiment analysis (Pang and Lee, 2004; Mullen and Collier, 2004; Wilson et al., 2005; Prabowo and Thelwall, 2009; Boldrini et al., 2009; Fernández et al., 2011). Following some examples of terms in this resource are described.

Example 1. Unigram `"bien"` (in English, *well*) has the following values:

$$\begin{aligned} pos(\text{“}bien\text{”}) &= 0.242 \\ neg(\text{“}bien\text{”}) &= 0 \\ obj\_measured(\text{“}bien\text{”}) &= 0.758 \\ neu(\text{“}bien\text{”}) &= 0.133 \\ real\_obj(\text{“}bien\text{”}) &= 0.075 \end{aligned}$$

Example 2. Instead, bigram `"bien vago"` (in English, `"rather vague"`) has these values:

$$\begin{aligned} pos(\text{“}bien\ vago\text{”}) &= 0 \\ neg(\text{“}bien\ vago\text{”}) &= 0.025 \\ obj\_measured(\text{“}bien\ vago\text{”}) &= 0.975 \\ neu(\text{“}bien\ vago\text{”}) &= 0 \\ real\_obj(\text{“}bien\ vago\text{”}) &= 0 \end{aligned}$$

As we can see in the examples, the word `"bien"` has suffered drastic changes regarding its scores. These facts indicate that using this method, in a statistical and autobalanced manner, we can obtain lexical evidences that are able to assume polarity scores depending of the context. It is important to remark that the reliability of this method will be increased as more tweets are explored, achieving a better balance on the polarities values.

4 Classifier II: Skipgram Scorer

Our second approach is similar to the first one, because it also creates a sentiment resource. To build this resource it is necessary to have a dataset of texts annotated with their polarity. This sentiment resource also consists on a lexicon that assigns a score to each term and each polarity. An important difference is that not only unigrams and bigrams are considered but also *skipgrams*. Skipgrams are a technique largely used in the field of speech processing, whereby n-grams are formed (bigrams, trigrams, etc.) but in addition to allowing adjacent sequences of words, it also allows tokens to be *skipped* (Guthrie et al., 2006). More specifically, in a *k-skip-n-gram*, n determines the number of terms, and k the maximum number of skips allowed.

In addition, the polarity scoring is different to the one used in the other approach, adding the scores of the skipgrams in the text taking into account for each one *i)* the number of skipped terms, *ii)* the number of occurrences, and *iii)* the proportion of occurrences in a specific polarity.

[6] http://www.cs.waikato.ac.nz/ml/weka/

4.1 Sentiment Resource

We preprocess each text in the dataset by removing accents and converting it to lower case. Then, each text is tokenised into terms, extracting only combinations of letters and numbers, in addition to Twitter users (starting with `@`) and hashtags (starting with `#`).

Afterwards, we obtain all the possible skipgrams from those terms by making combinations of adjacent terms and skipping some of them. The maximum number of terms in the skipgram is defined by the variable n and the maximum number of terms skipped is determined by the variable k. An example is shown in Figure 3.

Tweet:
"Conozco a alguien q es adicto al drama! Ja ja ja te suena d algo!"
↓
Skipgrams $(n = 2, k = 1)$**:**
conozco, conozco a, conozco alguien, a, a alguien, a q, alguien, alquien q, alguien es, q, q es, q adicto, es, es adicto, es al, adicto, adicto al, adicto drama, al, al drama, al ja, drama, drama ja, drama ja, ja, ja ja, ja ja, ja, ja ja, ja te, ja, ja te, ja suena, te, te suena, te d, suena, suena d, suena algo, d, d algo, algo

Figure 3: Example of skipgrams generation

In the next step, we calculate the scores for each skipgram. Each occurrence of a skipgram in each text is locally scored. This *local score* penalises skipgrams with a high number of skipped terms. The formula applied to calculated is shown in Equation 1, where s_t represents an occurrence of skipgram s in text t, $local(s_t)$ is the local score of the occurrence s_t, and the function *skipped* returns the number of skipped terms of the input skipgram occurrence.

$$local(s_t) = \frac{1}{skipped(s_t) + 1} \quad (1)$$

The *global score* of each single skipgram is calculated by adding the local scores of all occurrences of that skipgram. The formula that describes this method can be seen in Equation 2, where T represents the set of texts in the dataset, s_t represents an occurrence of skipgram s in text t, and $global(s)$ represents the global score of skipgram s. The *global polarity score* of each skipgram is similar to the previous score, but it only takes into account the texts with a specific polarity. The formula is presented in Equation 3, where p represents a specific polarity, T_p is the set of texts in the training corpus annotated with polarity p, s_t represents an occurrence of skipgram s in text t, and $global(s, p)$ represents the global score of skipgram s related to polarity p.

$$global(s) = \sum_{t \in T} \sum_{s_t \in t} local(s_t) \quad (2)$$

$$global(s, p) = \sum_{t \in T_p} \sum_{s_t \in t} local(s_t) \quad (3)$$

At the end of this process we have a list of skipgrams with a *global score* and a *global polarity score*: our second sentiment resource.

4.2 Features

The features for this approach are obtained from the previously created resource. Each text is preprocessed as mentioned previously, extracting its skipgrams. We obtain the global score and the global polarity score of each skipgram from the resource, and combine them to generate a single score, taking into account different factors:

- *Number of skipped terms.* Skipgrams with less skipped terms use to be more specific. The formula employed can be seen in Equation 4, where s_t represents an occurrence of skipgram s in text t, $f_{skipped}(s_t)$ is the factor of skipped terms of the occurrence s_t, and the function *skipped* returns the number of skipped terms of the input skipgram occurrence.

$$f_{skipped}(s_t) = \frac{1}{skipped(s_t) + 1} \quad (4)$$

- *Proportion of occurrences* in a text with a specific polarity. The bigger this proportion is, the more related the skipgram is related to that polarity. The formula proposed is described in Equation 5, where $f_{ratio}(s, p)$ is the proportion of occurrences of skipgram s related to polarity p.

$$f_{ratio}(s, p) = \frac{global(s, p)}{global(s)} \quad (5)$$

- *Number of occurrences* in a text with a specific polarity. Skipgrams that appear a high number of times related to a polarity are more relevant to that polarity. The formula is shown in Equation 6, where $f_{count}(s,p)$ is a count factor of occurrences of skipgram s related to polarity p, which increases its value as the number of occurrences is higher in a normalized way.

$$f_{count}(s,p) = 1 - \frac{1}{global(s,p)+1} \quad (6)$$

The score of an occurrence of a skipgram in a text for an specific polarity, is the product of all the factors described, using the formula in Equation 7.

$$\begin{aligned} score(s_t,p) &= f_{skipped}(s_t) \\ &\times f_{ratio}(s,p) \\ &\times f_{count}(s,p) \end{aligned} \quad (7)$$

The final score of the whole text for an specific polarity is calculated by adding the scores of all its skipgrams for that polarity, as shown in Equation 8.

$$score(t,p) = \sum_{s_t \in t} score(s_t,p) \quad (8)$$

Finally, to build a classifier, we use each polarity as a feature and each text as an instance. In the context of the competition, features are `P`, `P+`, `N`, `N+`, `NEU` and `NONE`. The value of a feature (polarity p) in an instance (text t) will be calculated using the function $score(t,p)$ (Equation 8). They will be used to create the polarity classifier employing a machine learning algorithm.

4.3 TASS Implementation

To create the sentiment resource we used only the dataset provided by the TASS 2013 competition organisers, composed by 7,219 tweets, which distribution is shown in Table 1. Using the features described, a classifier is created using the Weka default implementation of the *Support Vector Machines* (SVM) algorithm. Following some examples of terms in this resource are described.

Example 3. Skipgram `"bien"` (in English, `"well"`) has the following values:

$$\begin{aligned} global(\text{“}bien\text{”}, P) &= 16.0 \\ global(\text{“}bien\text{”}, P+) &= 55.0 \\ global(\text{“}bien\text{”}, N) &= 19.0 \\ global(\text{“}bien\text{”}, N+) &= 5.0 \\ global(\text{“}bien\text{”}, NEU) &= 6.0 \\ global(\text{“}bien\text{”}, NONE) &= 10.0 \end{aligned}$$

Example 4. Instead, skipgram `"bien mentira"` (in English, `"rather false"`) has these ones:

$$\begin{aligned} global(\text{“}bien\ mentira\text{”}, P) &= 0 \\ global(\text{“}bien\ mentira\text{”}, P+) &= 0 \\ global(\text{“}bien\ mentira\text{”}, N) &= 0 \\ global(\text{“}bien\ mentira\text{”}, N+) &= 0.33 \\ global(\text{“}bien\ mentira\text{”}, NEU) &= 0 \\ global(\text{“}bien\ mentira\text{”}, NONE) &= 0 \end{aligned}$$

Example 5. In addition, skipgram `"bien visto"` (in English, `"well seen"`) has these ones:

$$\begin{aligned} global(\text{“}bien\ visto\text{”}, P) &= 0 \\ global(\text{“}bien\ visto\text{”}, P+) &= 1.0 \\ global(\text{“}bien\ visto\text{”}, N) &= 0 \\ global(\text{“}bien\ visto\text{”}, N+) &= 0 \\ global(\text{“}bien\ visto\text{”}, NEU) &= 0 \\ global(\text{“}bien\ visto\text{”}, NONE) &= 0 \end{aligned}$$

As we can see in the examples, the word `"bien"` has also suffered changes regarding its scores, depending on the context. Without context, this word is more likely to have a strong positive polarity. But having the context into account, the polarity can vary to strong negative. Skipgrams are also useful to consider the context into the resource.

5 Evaluation

In order to assess the effectiveness of our classifiers, we performed a series of experiments over the provided datasets. The measures used are the traditional ones: *precision* (Pr) and *recall* (R). We do not use *accuracy* because it is not a good measure for text categorisation when using an imbalanced corpus (Yang and Liu, 1999). Instead, we also use the *F-score* (F1) because it represents a balance between precision and recall.

5.1 Train Evaluation

The first evaluation is performed over the training dataset using *10-fold cross validation* because of the small size of this corpus. The result of the evaluation of Classifier I (C1) and Classifier II (C2) is shown in Table 5.

The best results are obtained using the first approach RA-SR. The most probable

	Polarity	Pr	R	F1
C1	P	0.429	0.003	0.006
	P+	0.591	0.903	0.715
	N	0.516	**0.874**	0.649
	N+	0.625	0.028	0.053
	NEU	0.815	0.811	0.813
	NONE	**0.819**	0.867	**0.842**
	General	0.645	0.645	0.645
C2	P	0.160	0.018	0.032
	P+	0.366	0.088	0.140
	N	0.453	0.539	**0.492**
	N+	**0.487**	0.217	0.299
	NEU	0.319	0.279	0.297
	NONE	0.356	**0.696**	0.471
	General	0.384	0.384	0.384

Table 5: Results of the evaluation of our classifiers over the training dataset

reason for this fact is that this approach created a much bigger sentiment resource, so it has a much broader knowledge. An explanation for the low results of the second approach is that the training corpus is too small to obtain knowledge only from it.

As shown in Table 1, the corpora provided is highly imbalanced, so we can expect polarities with a lower number of texts to obtain worse results. However, there seems to be no direct relation between the number of texts and the results of each polarity in any of the systems, what can mean that our system and the corpora is robust against overtraining.

5.2 Test Evaluation

For the evaluation over the provided test dataset, we created a new classifier combining our approaches into a simple voting classifier. Each classifier returns a normalised value of certainty for each category. The new classifier will add those values to each certainty and choose the polarity with the higher value. For example, when a text is classified, if the first classifier returns a value of 0.25 for P and a value of 0.4 for NEU, and the second one returns a value of 0.3 for P+ and a value of 0.25 for P, the third classifier would have a value of 0.3 for P+, a value of 0.4 for NEU and a value of 0.5 ($= 0.25 + 0.25$) for P, so the final polarity for that text would be P.

These three classifiers were presented to the TASS competition. The evaluation over the provided test dataset was performed by the organisers from the results sent by the participants. In Table 6 we can see the top 10 systems with better results presented to the competition, where all our approaches are involved, and the combination of them (C3) reaches the first position. This fact suggests that our approaches are promising and encourages us to continue with the research and development of our systems.

	System	Pr	R	F1
1	**UA (C3)**	**0.616**	**0.616**	**0.616**
2	Elhuyar	0.601	0.601	0.601
3	Elhuyar	0.599	0.599	0.599
4	**UA (C2)**	**0.596**	**0.596**	**0.596**
5	UPV	0.576	0.576	0.576
6	UPV	0.574	0.574	0.574
7	UPV	0.573	0.573	0.573
8	CITIUS	0.558	0.558	0.558
9	Lys	0.553	0.553	0.553
10	**UA (C1)**	**0.552**	**0.552**	**0.552**
...	...	...	...	...

Table 6: Top 10 systems presented to the TASS 2013 competition for polarity classification with 6 levels

In Table 7 we can see the results of our systems in detail and divided by polarity. The best results are obtained by the combination of our two approaches. Combining two classifiers usually entails to worsen the best one, and improve the worst one of them, obtaining average results but, in this case, the combination resulted in a notable improvement. This fact can be due to both approaches share some similarities and they have been exploited.

Again, there seems to be no direct relation between the number of texts and the performance on each polarity, so we can deduce our approaches are still robust against overtraining. In all the approaches, the polarities with the lowest results are the neutral and positive ones. Detecting a neutral polarity can be very difficult compared to the other polarities: a neutral polarity can be due to the lack of polarity, but also when the positive and negative polarities are balanced. The positive polarity obtains also very low results, in contrast to the strong positive polarity, which is the polarity with better results in all the systems. A possible explanation for this is that the differences between different intensities of the positive polarity are so small that

	Polarity	Pr	R	F1
C1	P	0.019	0.001	0.001
	P+	**0.681**	0.604	**0.640**
	N	0.398	0.586	0.474
	N+	0.494	0.043	0.080
	NEU	0.103	0.131	0.115
	NONE	0.594	**0.657**	0.624
	General	0.552	0.552	0.552
C2	P	0.262	0.421	0.323
	P+	**0.739**	0.608	**0.667**
	N	0.527	0.510	0.518
	N+	0.585	0.441	0.503
	NEU	0.171	0.092	0.120
	NONE	0.574	**0.704**	0.633
	General	0.596	0.596	0.596
C3	P	0.358	0.263	0.304
	P+	**0.702**	**0.706**	**0.704**
	N	0.504	0.586	0.542
	N+	0.601	0.390	0.473
	NEU	0.156	0.108	0.128
	NONE	0.636	0.649	0.642
	General	0.616	0.616	0.616

Table 7: Results of the evaluation of our classifiers over the test dataset

classifiers tend to assign the most probable case. This is not the case of the negative polarity, where the differences seem to be bigger and classifiers can discriminate them with a better performance.

6 Conclusions and Future Work

In this paper, we presented our contribution for the *Task 1* (6 levels polarity classification) of the *TASS 2013* competition. This contribution consists on two different approaches: a modified version of a ranking algorithm (RA-SR) using bigrams, and new proposal using a skipgrams scorer. These approaches create sentiment lexicons able to retain the context of the terms.

Based on what we have presented, we developed a system that could solve the SA challenge with promising results. Our approaches appear in the top 10 best results of the systems presented to the competition, and the combination of them reaches the first position. This fact suggests that our approaches are promising and encourages us to continue with the research and development of our systems.

As future work, we propose to evaluate our approaches on different corpora and different domains, in order to check their robustness; deal with the neutral polarity finding more words to evaluate the obtained sentiment resource; and find the small difference between different intensities of polarities to improve the results of our SA task.

References

[Amigó et al.2012] Amigó, Enrique, Adolfo Corujo, Julio Gonzalo, Edgar Meij, and Maarten de Rijke. 2012. Overview of RepLab 2012: Evaluating Online Reputation Management Systems. In *CLEF (Online Working Notes/Labs/Workshop)*.

[Balahur et al.2010] Balahur, Alexandra, Ester Boldrini, Andrés Montoyo, and Patricio Martínez-Barco. 2010. The OpAL system at NTCIR 8 MOAT. In *Proceedings of NTCIR-8 Workshop Meeting*, pages 241–245.

[Boldrini et al.2009] Boldrini, Ester, Javier Fernández Martínez, José Manuel Gómez Soriano, Patricio Martínez Barco, et al. 2009. Machine learning techniques for automatic opinion detection in non-traditional textual genres.

[Esuli and Sebastiani2006] Esuli, Andrea and Fabrizio Sebastiani. 2006. Sentiwordnet: A publicly available lexical resource for opinion mining. In *Proceedings of LREC*, volume 6, pages 417–422.

[Fernández et al.2011] Fernández, Javi, Ester Boldrini, José Manuel Gómez, and Patricio Martínez-Barco. 2011. Evaluating EmotiBlog robustness for sentiment analysis tasks. In *Natural Language Processing and Information Systems*. Springer, pages 290–294.

[Guthrie et al.2006] Guthrie, David, Ben Allison, Wei Liu, Louise Guthrie, and Yorick Wilks. 2006. A closer look at skip-gram modelling. In *Proceedings of the 5th international Conference on Language Resources and Evaluation (LREC-2006)*, pages 1–4.

[Gutiérrez et al.2013a] Gutiérrez, Yoan, Andy González, Antonio Fernández Orquín, Andrés Montoyo, and Rafael Muñoz. 2013a. RA-SR: Using a ranking algorithm to automatically building resources for subjectivity analysis over

annotated corpora. *WASSA 2013*, page 94.

[Gutiérrez et al.2013b] Gutiérrez, Yoan, Andy González, Roger Pérez, José I Abreu, Antonio Fernández Orquín, Alejandro Mosquera, Andrés Montoyo, Rafael Muñoz, and Franc Camara. 2013b. UMCC_DLSI-(SA): Using a ranking algorithm and informal features to solve Sentiment Analysis in Twitter.

[Hall et al.2009] Hall, Mark, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. 2009. The WEKA data mining software: an update. *ACM SIGKDD Explorations Newsletter*, 11(1):10–18.

[Hu and Liu2004] Hu, Minqing and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177. ACM.

[Kozareva et al.2013] Kozareva, Zornitsa, Preslav Nakov, Alan Ritter, Sara Rosenthal, Veselin Stoyonov, and Theresa Wilson. 2013. Sentiment analysis in twitter. In *Proceedings of the 7th International Workshop on Semantic Evaluation. Association for Computation Linguistics.*

[Liu, Hu, and Cheng2005] Liu, Bing, Minqing Hu, and Junsheng Cheng. 2005. Opinion observer: analyzing and comparing opinions on the web. In *Proceedings of the 14th international conference on World Wide Web*, pages 342–351. ACM.

[Miller and Fellbaum1998] Miller, George and Christiane Fellbaum. 1998. WordNet: An electronic lexical database.

[Miller1993] Miller, George A. 1993. Five papers on WordNet. *Technical Report CLS-Rep-43, Cognitive Science Laboratory, Princeton University.*

[Mullen and Collier2004] Mullen, T. and N. Collier. 2004. Sentiment analysis using support vector machines with diverse information sources. In *Proceedings of EMNLP*, volume 4, pages 412–418.

[Padró et al.2012] Padró, Lluís, Miquel Collado, Samuel Reese, Marina Lloberes, Irene Castellón, et al. 2012. Freeling 2.1: Five years of open-source language processing tools.

[Pang and Lee2004] Pang, B. and L. Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, page 271. Association for Computational Linguistics.

[Pérez-Rosas, Banea, and Mihalcea2012] Pérez-Rosas, Verónica, Carmen Banea, and Rada Mihalcea. 2012. Learning Sentiment Lexicons in Spanish. In *LREC*, pages 3077–3081.

[Prabowo and Thelwall2009] Prabowo, R. and M. Thelwall. 2009. Sentiment analysis: A combined approach. *Journal of Informetrics*, 3(2):143–157.

[Sebastiani2002] Sebastiani, F. 2002. Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, 34(1):1–47.

[Strapparava and Valitutti2004] Strapparava, Carlo and Alessandro Valitutti. 2004. WordNet Affect: an Affective Extension of WordNet. In *LREC*, volume 4, pages 1083–1086.

[Villena Román et al.2013] Villena Román, Julio, Sara Lana Serrano, Eugenio Martínez Cámara, and José Carlos González Cristóbal. 2013. TASS-Workshop on sentiment analysis at sepln.

[Wiebe, Wilson, and Cardie2005] Wiebe, Janyce, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language resources and evaluation*, 39(2-3):165–210.

[Wilson et al.2005] Wilson, T., P. Hoffmann, S. Somasundaran, J. Kessler, J. Wiebe, Y. Choi, C. Cardie, E. Riloff, and S. Patwardhan. 2005. OpinionFinder: A system for subjectivity analysis. In *Proceedings of HLT/EMNLP on Interactive Demonstrations*, pages 34–35. Association for Computational Linguistics.

[Yang and Liu1999] Yang, Y. and X. Liu. 1999. A re-examination of text categorization methods. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 42–49. ACM.