Studying the Scope of Negation for Spanish Sentiment Analysis on Twitter

Salud María Jiménez-Zafra, M. Teresa Martín-Valdivia, Eugenio Martínez-Cámara, and L. Alfonso Ureña-López

Abstract—Polarity classification is a well-known Sentiment Analysis task. However, most research has been oriented towards developing supervised or unsupervised systems without paying much attention to certain linguistic phenomena such as negation. In this paper we focus on this specific issue in order to demonstrate that dealing with negation can improve the final system. Although we can find some studies of negation detection, most of them deal with English documents. On the contrary, our study is focused on the scope of negation in Spanish Sentiment Analysis. Thus, we have built an unsupervised polarity classification system based on integrating external knowledge. In order to evaluate the influence of negation we have implemented a specific module for negation detection by applying several rules. The system has been tested considering and without considering negation, using a corpus of tweets written in Spanish. The results obtained reveal that the treatment of negation can greatly improve the accuracy of the final system. Moreover, we have carried out a comprehensive statistical study in order to demonstrate our approach. To the best of our knowledge, this is the first work which statistically demonstrates that taking into account negation significantly improves the polarity classification of Spanish tweets.

Index Terms—Negation scope, sentiment analysis, twitter, spanish opinion mining, polarity classification, lexicon based system, statistical analysis

1 Introduction

Sentiment Analysis (SA) is a challenging area related to Text Mining and Natural Language Processing (NLP) that combines several techniques in order to examine subjectivity in textual documents computationally. There are several issues related to SA such as subjectivity detection, opinion extraction, irony detection and so on. In this paper we focus on one of these tasks, polarity classification. This task aims to determine the overall sentiment-orientation (positive, negative or neutral) of the opinions contained within a given document.

Although at the beginning, most research has been oriented towards analyzing the sentiments in forums or web sites like Amazon or Epinion, the use of social networks as a huge source of data for SA is becoming more and more important. Specifically, micro-bloggings such as Twitter are being used to measure voting intention, consumer opinions and people's moods. In the last years, the number of scientific papers combining SA and Twitter has increased exponentially. However, most of this research is oriented to documents/tweets written in English, perhaps due to the novelty of the task and the lack of resources in other languages. Nonetheless people increasingly comment on their experiences, opinions, and points of view not only in English but in many other languages. Consequently, the management and study of subjectivity and SA in languages other than English is a growing need. The work presented herein is focused on polarity classification of Spanish tweets.

On the other hand, although polarity classification is the most widely studied task in SA, several challenges still remain open and are attracting the attention of researchers. One of these is the treatment of some linguistic phenomena such as irony, metaphors or negation. In this paper we focus on the treatment of negation. Actually, our main goal is to statistically demonstrate whether the detection and integration of negation in polarity classifier of Spanish tweets can improve the accuracy of the final system. To this end we first study the different cues that work as triggers of negativity. Then we define several rules in order to detect the scope of negation. Finally, we use our unsupervised lexicon based system to classify the polarity of a tweet. We demonstrate, by carrying out a statistical significance study, that the detection of negation and the application of some heuristic rules can significantly improve the final system.

The rest of the paper is organized as follows. The next section outlines existing research that has served as the basis of our work. Then the main resources used are described. Later, we introduce a section to study the scope of negation. Section 5 describes the architecture of the proposed unsupervised approach. Next, the set of experiments that we have carried out are specified and analyzed. Moreover, we include a specific section to statistically demonstrate the validity of our approach. Finally, we conclude our study and present future directions for research.

• *The authors are with the Department of Computer Science, Universidad de Jaén, Campus Las Lagunillas, Jaén E-23071, Spain. E-mail: {sjzafra, maite, emcamara, laurena}@ujaen.es.*

Manuscript received 4 Apr. 2016; revised 5 Dec. 2016; accepted 9 Apr. 2017. Date of publication 11 Apr. 2017; date of current version 7 Mar. 2019. Recommended for acceptance by F. Pianesi. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below. Digital Object Identifier no. 10.1109/TAFFC.2017.2693968

2 Related Work

Our main goal is to demonstrate the usefulness of considering negation in a polarity classification system over a corpus of tweets. Thus, we will present some studies that have been the basis of our work. First we talk about how Twitter is considered one of the main sources of opinions that can be exploited by the SA community, paying special attention to papers dealing with negation and Spanish. Then, we describe some other unsupervised systems based on lexicons because our model uses this kind of linguistic resource. Finally, a review of different papers studying negation is presented.

2.1 Sentiment Analysis on Twitter

The research community of SA was one of the first to become aware of the potential of Twitter as a great source of information to extract and generate knowledge from the data that users post [1]. Perhaps the first work related to the study of user opinions in this social network was presented by [2]. The authors developed a supervised system with the aim of analyzing the most suitable lexical features to represent a tweet and the most acceptable machine-learning algorithm to identify the polarity of the tweet. After this study a wide range of methods for SA on Twitter have been published, describing systems with different features and methodologies including supervised systems [3], [4], unsupervised approaches [5] and hybrid methods [6], [7].

However, few papers explicitly focus on negation or Spanish texts and only a few studies take into account features related to these issues. In [8] the performance of the unsupervised version of the algorithm SentiStrength is evaluated. The system combines several opinion lexicons of words and idioms. The system also incorporates a straightforward method for detecting the scope of negation cues that does not invert the polarity of a word, but makes the word neutral. The authors conclude that their lexicon-based proposal is suitable for polarity classification in Twitter. In [9] a supervised polarity classification system for English tweets is described. The system follows the same approach for defining negation as [10], that is, all subsequent words to a negative particle are considered as negated words. In [11] a novel method is shown for calculating the polarity of a given tweet following an unsupervised approach. Montejo-Ráez makes use of the semantic resource WeFeelFine [12], which is a huge database of sentences related to feelings and emotions. The sentiment system is similar to a search engine, where each tweet is treated as a query and the system returns the hundred emotions most similar to the given tweet. The final polarity is obtained by a weighted sum of the polar score of those hundred most similar emotions. The same approach is applied to Spanish reviews with promising results in [13]. Finally, in [14] an experimental study exploring the lexical and syntactic information in Spanish tweets in order to improve a polarity classification system is presented. The architecture is composed of different modules including a methodology to identify the scope of negation using a few negation terms and considering only four dependency-based rules. The experimental framework used to evaluate the systems is the corpus supplied by the TASS2013 organizers [15]. Our work is very close to this approach, but we consider more rules and more particles to detect the scope of negation. In addition, in this paper we carry out a statistical study in order to demonstrate the benefits obtained from applying these rules.

2.2 Lexicon Based Systems

In this paper we present a system that follows a lexicon-based strategy, so in the following lines we expound some papers related to this approach. For example, in [16] a lexicon-based method is used to take advantage of WordNet in building a lexicon of opinion bearing words. The polarity lexicon is used in conjunction with a sentiment lexicon of hashtags and a module for the identification of the scope of negation in order to develop a Twitter SA system in the political domain. When a lexical-based method is selected to build a polarity classifier, the building of a new list of opinion bearing words is not mandatory, the use of an existing lexicon is possible. This is the case of the paper [17], where the authors use the opinion lexicon General Inquirer to classify both subjectivity and polarity.

In order to follow a lexicon based method, a list of opinion bearing words is needed. Three Spanish opinion lexicons are the most well-known by the SA research community. In 2012 [18] was published, wherein the authors describe a framework that generates sentiment lexicons in a target language by using manually and automatically annotated English resources. The target language in the paper is Spanish, so the authors built two Spanish opinion lexicons, one from a manually labelled English opinion lexicon and another from an automatically labelled English opinion lexicon. Despite its recent publication, the opinion lexicon of Perez-Rosas et al. is being used in some studies such as [19] and [20]. Another interesting lexicon is described in [21], where the authors present a dictionary marked with probabilities to express one of the six basic emotions. The dictionary, which is known as the Spanish Emotion Lexicon (SEL), contains 2,036 words. Due to the fact that each word of the lexicon is labelled with an emotion and not with a polar label, the lexicon is less used by the SA research community. The third opinion lexicon is iSOL, which is described in [22] and has also been used successfully in [23] and [24]. iSOL is the lexicon used for determining the polarity of the system presented in this paper, and it will be described in the next section.

2.3 Negation and Sentiment Analysis

Regarding the treatment of negation, most research has focused on opinions written in English. One of the first approaches was proposed in [25] using a simple method that adds "NOT" to the terms of the sentence that appear next to negative terms, such as "no" or "don't". In [10] the same approach is followed, but they assume that the negation cues ("not", "isn't", "didn't", etc.) affect all the terms from the cue to the end of the sentence. The authors carry out different experiments with and without negation using machine learning algorithms. However, the results show no significant differences considering negation or not. In [26] not only is negation considered but they also study intensifiers and diminishers, introducing the new concept "Contextual valence shifters". In [27] a similar methodology is used where negations are used to reverse the semantic polarity of a particular term, while intensifiers and diminishers are used to increase and decrease, respectively, the degree to which a term is positive or negative. In addition, in [28] an unsupervised model is proposed based on a fixed window of 4 words to determine negation scope. Other current researchers are developing rule-based systems using syntactic dependence trees [29] or applying more complex calculations in order to obtain polarity in opinions [30].

All these studies deal with English texts. There are even some good surveys about the study of negation as a linguistic phenomenon [31] and concerning SA [32]. However, for Spanish SA it is very difficult to find research considering negation as a feature. In [33] the same approach as the one used for English is applied, but adapted to Spanish. Thus, using their SO-CAL tool [34] they evaluate several negation cues and calculate the polarity values depending on different features related to the terms and the grammatical category. Finally, in [35] the syntactic structure of the text is considered, showing an improvement over the systems that only use lexical features. Their recent work [14] shows some interesting results over the Spanish corpus of tweets supplied in the TASS2013 workshop [15]. However, they do not make any analysis of the gain obtained using negation individually, and so it is not possible to determine which is the module responsible for the improvement obtained.

3 Resources

Increasingly, linguistic resources are becoming key players in NLP systems because they are the source of knowledge needed by NLP systems to achieve their primary objective, which is the understanding of natural language. Furthermore, linguistic resources are necessary due to the fact that the performance and the quality of NLP systems have to be assessed. Therefore, two kinds of linguistic resources can be distinguished: The first ones are mainly employed as an essential element to building NLP systems, and the second ones are tools for evaluating such systems. The present paper describes a study in which the two sorts of linguistic resources are used with the aim of showing the importance of taking into consideration negation in the context of polarity classification on Twitter in Spanish.

The polarity classification system developed for the study follows a lexicon-based approach, so some sets of sentiment-bearing expressions have been employed. Specifically, we consider a list of opinion words, a set of emoticons separated by the sentiments represented by them, and a list of hashtags that express sentiment.

In addition, a corpus of Spanish tweets is necessary for the assessment. Currently, two corpora of Spanish tweets are available for the research community. The first one is the corpus used in the TASS workshop [15], and the second one is the Corpus Of Spanish Tweets COST[1] [36]. In this paper we have chosen the TASS corpus for several reasons. First, the TASS corpus is broadly known by the Spanish research community, due mainly to the fact that it has been used in the previous four editions of the TASS workshop; the TASS corpus, which has about 68,000 tweets, contains considerably more tweets than the COST corpus, which is only composed of 34,634 tweets; and finally the TASS corpus was labelled following a semi-automatic process while the COST was labelled following a noisy label approach, which is similar to the one employed in [2].

3.1 iSOL Lexicon

Although Spanish SA is attracting more and more researchers, the number of opinion lexicons is scarce compared to the ones available in English. For English SA we can find several resources such as the opinion lexicon compiled by Bing Liu [37], the MPQA lexicon [28], General Inquirer [38], SentiWordNet [39] and so on.

However, for Spanish the number of resources is limited. In this paper, we have used the iSOL lexicon because it has been successfully applied in other studies. iSOL is a Spanish lexicon composed of 8,135 opinion words (2,509 positive words and 5,626 negative words). This resource was created taking as a basis the list of opinion words compiled by Bing Liu, which was translated into Spanish. Subsequently, the translated version of the list was manually reviewed and it was completed with more Spanish terms in order to obtain a more representative list of Spanish opinion words. All the details of the compilation process of iSOL can be found thoroughly described in [22]. The evaluation of iSOL demonstrates its validity for sentiment analysis in Spanish.

3.2 Hashtags, Emoticons and Laughs

The language used in Twitter has two special elements that are constantly typed by users, mentions and hashtags. A mention is the explicit reference that a user makes to another through writing the username preceded by the @ symbol. A hashtag is a string preceded by the hash key (#), and it is usually employed in order to identify the main topic, the sentiment or the semantic orientation of the tweet. Thus, taking into consideration hashtags in the process of polarity classification of tweets in Spanish could be a good idea. In [40] the effect of hashtagging emotions such as joy, sadness, anger and surprise in order to express the general emotion or sentiment in a tweet is studied. In a later paper [9], the authors describe the compilation of a lexicon of opinion using hashtags in English. To our knowledge a lexicon of Spanish opinion hashtags is not available, so the compilation of a Spanish opinion hashtag lexicon was undertaken. For this, we used a seed of positive hashtags (#*bueno* (#good), #*bien* (#well), #*positivo* (#positive), #*fantastico* (#great), #*excelente* (#excellent), etc.) and another of negative hashtags (#*malo* (#bad), #*mal* (#bad), #*terrible* (#terrible), #*negativo* (#negative), #*horrible* (#horrible), etc.) and retrieved all the tweets that had any of the seed words for three days. Then, we extracted all the hashtags present in those tweets and classified them as positive or negative depending on whether they appeared in the same tweet of a positive or negative seed. Finally, we manually reviewed these hashtags in order to obtain the final lists. In this way, the hashtags lexicon[2] was compiled and it is composed of 172 positive and 127 negative hashtags.

Emoticons are other indicators of polarity that should be taken into account. In [41] it was shown that when the author of an electronic communication uses an emoticon,

1. http://sinai.ujaen.es/cost

2. http://sinai.ujaen.es/hashtags-sp

TABLE 1
Rules for Identifying the Scope of Negation Cues

Cue	Rule for scope identification
no (not), *tampoco* (neither), *nadie* (nobody), *jamás* (never), *ninguno* (none)	Parent node and the tree formed by the brother of the right, included
ni (nor), *sin* (without)	All children and all trees formed by them until reaching leaf nodes
nada (nothing), *nunca* (never)	Parent node

he/she is effectively marking up the text with an emotional state. In [2] emoticons are used to build one of the first corpus of tweets for SA. In [36] emoticons have also been used to compile a corpus of positive and negative tweets written in Spanish. According to the emotions itemized in Wikipedia,[3] two lists of emoticons were generated:[4] one of them with 70 positive emoticons and another one with 46 negative emoticons.

Laughs are another element frequently used in Twitter. For identifying them we have defined a regular expression with the main forms of writing laughs in Spanish and variants thereof: jajaja, jaaajajaj, jijiji, jijiij, lol, loool, etc.

3.3 The TASS Corpus

In order to evaluate our proposal we have used a corpus widely known by the Spanish SA research community, called General Corpus of TASS[5] [15]. It was published for the first time in 2012 and since then it has been used in all the subsequent editions of the workshop on SA at SEPLN (2013, 2014, 2015 and 2016), so up until now it is the main corpus of Spanish tweets tagged for SA. The corpus contains over 68,000 tweets gathered between November 2011 and March 2012. The tweets were written in Spanish by about 150 well-known personalities and celebrities of the world of politics, economy, communication, mass media and culture. The corpus is divided into two sets: training (10 percent) and test (90 percent), so the training set is composed of 7,219 tweets and the test one is formed by 60,017 tweets. Each tweet in both sets is tagged with its global polarity, indicating whether the text expresses a positive, negative or neutral sentiment, or no sentiment at all. Five levels have been defined: strong positive (P+), positive (P), neutral (NEU), negative (N), strong negative (N+) and one additional no sentiment tag (NONE).

We consider the TASS corpus has become a benchmark for Spanish SA on Twitter. Thus, we think it is a good choice for our experiments. Because our system is completely unsupervised and does not require training data, only the test set of the TASS corpus was taken into consideration for the assessment of the proposal. In addition, we neglected the tweets tagged with NONE class and only considered Positive, Negative and Neutral classes. Thus, original strong positive (P+) and positive (P) tweets are grouped into one unique positive class (P). Alike, strong negative (N+) and negative (N) are considered as negative class (N). After all this processing, the final set of tweets used for the assessment is composed of 22,233 positive tweets, 1,305 tweets labelled as neutral, and 15,844 negative tweets, which is a total of 39,381 tweets.

4 Negation Scope Identification

Negation is an important feature of language that requires a special treatment in the field of NLP and specifically in SA. It is considered a challenging task because it is a linguistic phenomenon that has not been studied enough, especially in Spanish.

The present paper is oriented towards the study of this challenge for SA: identification of the scope of negation in Spanish texts. Our main goal is to demonstrate whether by taking into account negation we can improve the polarity classification of Spanish tweets. We think that a correct identification of the negation scope could help in the polarity classification of a text because a negative opinion can be expressed using positive words negated (e.g., *No fue una buena idea asistir al concierto* (It was not a good idea to go to the concert)) or, by contrast, a positive opinion can be expressed from the negation of negative words (e.g., "*La actuación no fue un desastre como se esperaba*"/ The performance was not a disaster as expected).

As a first approach to this phenomenon, we propose a set of rules based on dependency trees for identifying the scope of some negation cues. In particular, we have studied the most important according to La Real Academia Española (Royal Spanish Academy) [42]: *no* (not), *tampoco* (neither), *nadie* (nobody), *jamás* (never), *ni* (nor), *sin* (without), *nada* (nothing), *nunca* (never) and *ninguno* (none). For each negation cue, a rule for determining its scope was defined. For this, we analyzed the dependency trees of diverse sentences extracted from different websites in which some of the cues considered appear. To build the dependency trees we used the dependency parser of Freeling [43], which generates the dependency tree of a sentence based on its syntactic structure. Freeling[6] [44] is an open-source language-analysis toolkit that is available for several languages, including Spanish. After the study of these trees, we realized that it is possible to generalize the treatment of these negation cues in 3 rules (Table 1). We analyzed, on average, ten dependency trees per negation cue. The dependency trees produced by Freeling were always coherent with the rules, so we decided to continue research and apply them to Spanish Twitter SA. Although we think that the use of a specialized parser is better for the processing of tweets, we also support the idea that while specialized parsers are not available, a standard parser can be used. To the best of our knowledge, there is no specific parser for Spanish tweets so the NLP tool most used for Spanish (Freeling) was chosen. Moreover, due to the fact that tweets are informal texts, we apply a spelling checker in order to keep the number of errors as low as possible and to make the dependency parser work successfully (see Section 5).

In order to clarify the rules that have been defined, an example of the applications of each rule is shown in Figs. 1, 3 and 2. Each figure represents the dependency tree related

3. http://es.wikipedia.org/wiki/Anexo:Emoticonos
4. http://sinai.ujaen.es/emoti-sp
5. http://www.daedalus.es/TASS2014/tass2014.php#corpus
6. http://nlp.lsi.upc.edu/freeling/

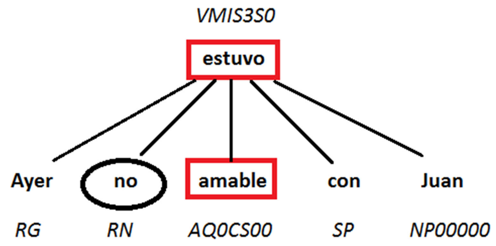

Fig. 1. Dependency tree of the negative word *no* (not). Tweet: *Ayer no estuvo amable con...* (Yesterday he was not kind to...).

to a tweet in which the negation cue is represented with an ellipse and its scope is marked with a box.

The integration of these rules in a polarity classification system allows us to tag the words that are in the scope of any negation cue, with the aim of taking this into account when the polarity of a tweet is determined. For example, in the tweet *Han actuado sin defensa ni garantías para los usuarios.* (They have acted without defense nor guarantees for the users.) (Fig. 2), the system will detect that there are two negative particles in the text, *sin* (without) and *ni* (nor), and for each one it will determine its scope using the rules defined. In this case, both particles affect all children nodes and all trees formed by them until reaching leaf nodes. Therefore, the words *defensa* (defense) and *garantías* (guarantees) will be tagged as negated words in order to modify their polarity value.

5 System Architecture

As we have mentioned earlier, the aim of this study is to demonstrate that taking into account negation is useful in a polarity classification system of tweets. To verify this assertion, we propose an unsupervised lexicon based system made up of different components. The main contribution of this system is the development of a normalization module that corrects misspelled words, another that detects the presence of a negation cue in a tweet and determines its scope using the rules defined, and the compilation of a Spanish opinion hashtags lexicon. The approach used for determining the polarity of a tweet is straightforward because our goal is not focused on demonstrating that our system is a good polarity classifier but showing that treatment of negation is useful in such systems. The processing

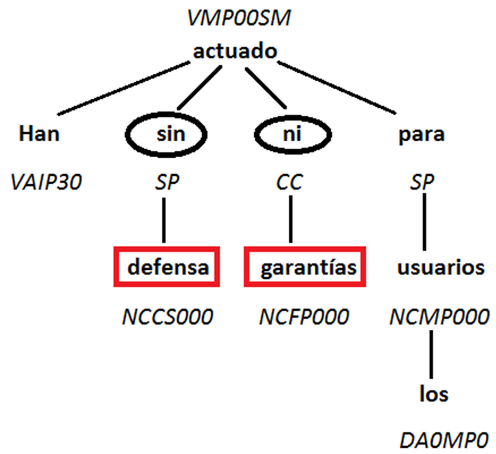

Fig. 2. Dependency tree of the negative words *sin* (without) and *ni* (nor). Tweet: *Han actuado sin defensa ni garantías para los usuarios* (They have acted without defense nor guarantees for the users).

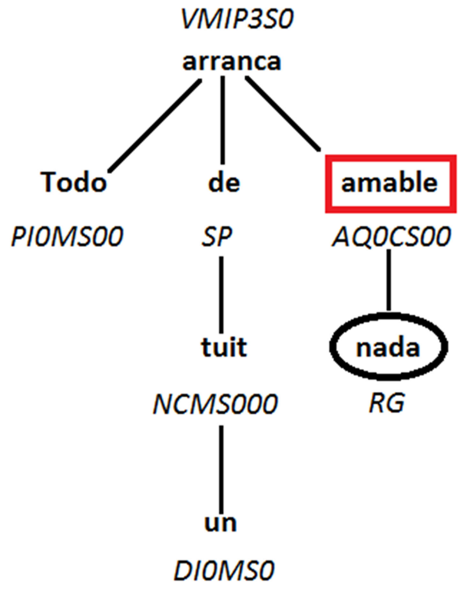

Fig. 3. Dependency tree of the negative word *nada* (nothing). Tweet: *Todo arranca de un tuit nada amable* (It all starts with an unkind tweet).

of each tweet to obtain a final polarity classification can be summarized in five steps:

1) Tokenize the tweet.
2) Correct misspelled words.
3) Determine the part of speech of each word and the lemma of each verb.
4) Detect the presence of negation cues and identify the scope of each of them using the rules defined.
5) Obtain the polarity of the tweet.

The process outlined is shown in Fig. 4. Below, a detailed explanation of all elements is shown with the sample tweet *Todo arranca de un tweet nada amaaable. #maldad =(* (It all starts with an unkind tweet. #wickedness = ().

1) *Tokenization:* In order to process the text in the tweet, sentence splitting and word tokenization have to be performed. For this, the Freeling splitter and an adapted version to the Spanish language of the Christopher Potts' tokenizer[7] were used. The tokenizer developed takes into account all special features of the language used in Spanish tweets: emoticons, urls, mentions, hashtags, dates, multiwords, etc. Below, the tokens that the system would identify in the sample tweet are shown in square brackets:

   [Todo][arranca][de][un][tweet][nada][amaaable][.][#maldad][= (]

2) *Normalization:* After the identification of the tokens, the next step is to perform a normalization process in order to correct all misspelled words and to mark the tokens that have repeated letters. We mark the tokens that have repeated letters to consider their intensity when we calculate the overall sentiment of the tweet. The two reasons for performing

7. http://sentiment.christopherpotts.net/tokenizing.html#sentiment

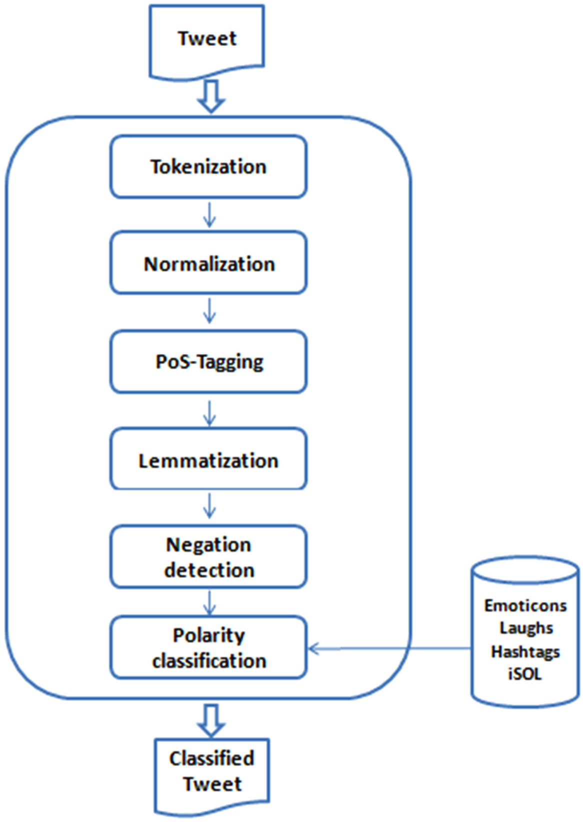

Fig. 4. Architecture of the polarity classification system.

normalization to correct spelling errors are, first, that our system needs to build the syntactic tree of each tweet, so if there are fewer misspellings in the text the dependency parser will be more likely to be successful. The second reason is that the system is based on the use of the lexical resource iSOL which is a list of words, most of them well written. The spelling corrector of Peter Norvig[8] has been modified with the aim of correcting misspellings in Spanish texts. This spelling corrector only needs to work a large corpus in the target language. In our case, the target language is Spanish, so we have to compile a representative corpus of Spanish. This large corpus is composed of a list of Spanish lemmas, a list of Spanish verb conjugations and a list of Spanish names and surnames. All the lists were compiled by Ismael Olea.[9] The initial lists were complemented by the list of words of the corpus CREA,[10] which was compiled by La Real Academia Española (Royal Spanish Academy). Normalization of the sample tweet would correct the token `[amaaable]` and would also mark it as a token with repeated letters:

[Todo][arranca][de][un][tweet][nada]
[**amable**][.][#maldad][= (]
Repeated letters

8. http://norvig.com/spell-correct.html
9. http://olea.org/proyectos/lemarios/
10. Royal Spanish Academy (http://www.rae.es): Data Bank (CREA) online. Current Spanish Benchmark Corpus. http://corpus.rae.es/creanet.html

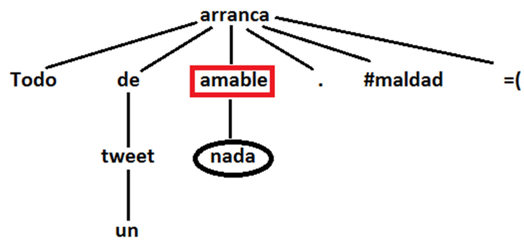

Fig. 5. Dependency tree of the tweet: *Todo arranca de un tweet nada amable. #maldad =(* (It all starts with an unkind tweet. #wickedness = (.).

3) *PoS-Tagging and Lemmatization:* The third step is to learn the PoS-tag of each token in order to obtain the lemma of each verb, because iSOL does not have all the verbal forms of polar verbs, it only has the lemma of each one. Therefore, we used the Part-of-Speech tagger module of Freeling. This resource has two different modules for performing PoS tagging [44]. The first one is the hmm tagger which is a classical trigram Markovian tagger [45] and the second one, named the relax tagger, is a hybrid system capable of integrating statistical and hand-coded knowledge [46]. We used the hmm tagger because it is faster than the relax tagger. In the case of the sample tweet, the system would tag each token with its pertinent part of speech and would obtain the lemma of the token `[arranca]` because it is a verbal form.

[Todo][**arrancar**][de][un][tweet][nada]
[amable][.][#maldad][= (]
Repeated letters

4) *Negation detection:* This module, in the first place, detects whether the tweet has any negation cue and if so, it determines the scope of each cue with the set of syntactic rules that has been defined (Table 1). In this way, if a tweet has a negation cue the system will generate its dependency parser and will mark each word affected by negation as "negated" by "name_of_the_cue", in order to take this into account when the semantic orientation of the tweet is calculated. In the sample tweet there is a negation cue, the token `[nada]`. In this case, the system would generate the dependency tree of the tweet (Fig. 5) and would mark as negated by `[nada]` the token `[amable]` that is in its scope according to the rule defined.

[Todo][arrancar][de][un][tweet][**nada**]
[**amable**][.][#maldad][= (]
Repeated letters
Negated by nada

5) *Polarity classification:* The last step is to determine the polarity of the tweet. For this purpose, a polarity classifier that takes into account the presence of emoticons, hashtags, expressions of laughing and negation was developed. This component uses the resources described in Section 3: the bag of words of emoticons tagged as positives and negatives, the bag of hashtags and iSOL lexicon.

For each tweet the classifier determines its positivity and negativity value. Thus, if a token is in the bag of positive/negative emoticons, a polarity value of 2 is added to its positivity/negativity value. If it detects that a token is an expression of laughing, the positivity value is increased by 2. In the other case, if the token is in the bag of positive/negative hashtags, the counter of positivity/negativity is increased by 2. Finally, if the token is in the iSOL positive/negative list, a polarity value of 1 is added to the positivity/negativity counter and if it also has repeated letters the value is increased by 1. If the token is negated its polarity is reversed (positive → negative, negative → positive). Using these values, the system is able to classify the tweet in one of the 3 defined classes following the equation

$$polarity(tweet) = \begin{cases} P & \text{if } pv > nv \\ NEU & \text{if } pv = nv \\ N & \text{if } pv < nv \end{cases}, \quad (1)$$

where pv and nv are the positivity and negativity value of the tweet respectively.

According to this approach, the sample tweet would be classified as negative because its positivity value would be 0 and its negativity value would be 7. Below, these values are explained with details (both values are initialized to 0, $nv = 0$, $pv = 0$):

- Negativity value:
  - `Arrancar` is a verb that belongs to the list of negative words of the iSOL lexicon ($nv + 1 = 1$).
  - `Amable` is an adjective that belongs to the list of positive words of iSOL, but it is tagged as negated token because it is in the scope of the negative particle *nada*. So its polarity value is reversed (positive → negative) ($nv + 1 = 2$) and it also has repeated letters ($nv + 1 = 3$).
  - `#maldad` is a negative hashtag ($nv + 2 = 5$).
  - `=(` is a negative emoticon ($nv + 2 = 7$).
- Positivity value:
  - As we have mentioned before, `Amable` is a positive adjective, but it is in the scope of the negative particle *nada* so its polarity value is reversed. Therefore, the positivity value remains 0 ($pv = 0$).

[`Todo`][`arrancar`](−1)[`de`][`un`][`tweet`]{[`nada`][`amable`](+2)}(−2)[.][`#maldad`](−2)[= (](−2)
Note : "amable" adds two negative points because it is a positive opinion word with repeated letters and it is negated by *nada* .

6 Experiments and Results

In this section we first present the measures applied in order to evaluate our approach, then we describe the different experiments carried out, and finally we expound the results obtained.

In order to evaluate our system we calculated the usual measures: Precision (P), Recall (R), F-score (F1) and Accuracy (Acc)

$$P = \frac{TP}{TP + FP}, \quad (2)$$

$$R = \frac{TP}{TP + FN}, \quad (3)$$

$$F1 = \frac{2PR}{P + R}, \quad (4)$$

$$Acc = \frac{TP + TN}{TP + TN + FP + FN}, \quad (5)$$

where TP (True Positives) are those assessments where the system and human experts agree on a label assignment, FP (False Positives) are those labels assigned by the system that do not agree with the expert assignment, FN (False Negatives) are those labels that the system failed to assign as they were given by the human expert, and TN (True Negatives) are those non-assigned labels that were also discarded by the expert. The Precision tells us how well the labels are assigned by our system (the fraction of assigned labels that are correct). The Recall measures the fraction of the experts labels found by the system. Finally F1 combine both Precision and Recall, while Accuracy takes into account all the correct results including TN [47]. For ease of comparison, we summarize the F1 scores over the different categories (positive, negative and neutral) using the Macro-averaged F1

$$Macro - F1 = \frac{1}{|c|} \sum_{i=1}^{|c|} \frac{2P_iR_i}{P_i + R_i}. \quad (6)$$

In the same way, we can obtain the Macro-Recall and Macro-Precision as follows:

$$Macro - R = \frac{1}{|c|} \sum_{i=1}^{|c|} \frac{TP_i}{TP_i + FN_i}, \quad (7)$$

$$Macro - P = \frac{1}{|c|} \sum_{i=1}^{|c|} \frac{TP_i}{TP_i + FP_i}. \quad (8)$$

6.1 Experiment Design

Our goal is to study whether the application of certain rules for detecting the scope of negation provides benefits in the polarity classification of Spanish tweets. Moreover, we compare the rules-based approach that we propose with the most widely used model to address negation in English Twitter SA [48]. In order to explore the contribution of this linguistic phenomenon we carried out the following experiments:

- Without negation (Baseline = BS): using the system described, but without taking into account negation.
- With a baseline for negation (BSN): using the BS system with the negation approach proposed in [48] that considers as scope of negation all terms from a negation cue to the next of the following punctuation marks: ".", ":", ";", "!", "?". This method is the most widely used in order to determine the scope of negation in English Twitter SA.

TABLE 2
Tweets Used in the Experiments

	Tweets	Percentage
Total	39,382	100%
NegCue	8,604	22%
RuleAffect	2,326	6%

- With Negation Rules (NR): using the BS system described in the paper, but including the module that detects the presence of a negation cue in a text and determines its scope using the syntactic rules defined.

The experiments were conducted on the tweets of the TASS corpus considering three cases:

- *Total*: all the tweets of the corpus tagged as positive (P), negative (N) or neutral (NEU) are considered. The total set is composed of 39,382 tweets.
- *NegCue*: we only take into account the tweets of the corpus that have any of the negation cues studied in this paper. The experiment is carried out over a set of 8,604 tweets.
- *RuleAffect*: we only consider the tweets that contain some of the negation cues studied and also polar tokens which are in the scope of these particles, i.e., opinion words affected by negation. The total number of tweets is 2,326

$$Total \supset NegCue \supset RuleAffect.$$

As we can see, *RuleAffect* is a subset of *NegCue* and *NegCue* is a subset of *Total*. The reason why the dataset has been reduced to carry out experiments with two subsets is because most of the tweets in the corpus do not have negation cues (Table 2), and so in order to determine whether negation improves the polarity classification we need to compare the subsets with and without negation cues.

Specifically, in order to evaluate the rules defined we should consider the tweets with polar tokens affected by negation because these are the tweets in which the rules have actually been applied. The rules are based on reversing the polarity of the words that are in the scope of negation, but if the lexicon used does not detect a polar token the rules can not be applied. The aim of this study is not to present the best polarity classification system but to check whether the rules that we have introduced can improve the classification of Spanish tweets.

In order to clarify the difference between these 3 datasets, we show below an example of the type of tweets present in each one. Below we can see a sample of a tweet that belongs to the *Total* set, but is not included in any other subsets (*NegCue*, *RuleAffect*) because it does not have any negation cue.

> *Gonzalo Altozano tras la presentación de su libro 101 españoles y Dios. Divertido, emocionante y brillante.* (Gonzalo Altozano after the presentation of his book 101 Spaniards and God. Fun, exciting and brilliant.)

Subsequently, a tweet of the *NegCue* subset is shown. This sample is part of the *Total* set, but it does not belong to the *RuleAffect* subset because the token *exacerbar* is not identified as a polar token by the iSOL lexicon, and consequently the rule of the negation cue *no* cannot be applied.

> *RT @usuario : Yo cuidaría como gestionar todo este panorama para no exacerbar: Nos jugamos demasiado...* (RT @user: I would look after how to manage this panorama to avoid exacerbating. We have a lot at stake...)

Finally, a tweet of the *RuleAffect* subset is presented. In this case, the term *amable* is in the scope of the negation cue *nada* and it is identified as an opinion word by the iSOL lexicon, so the rule can be applied and evaluated. As we can see, this subset only has tweets with negation cues and polar tokens affected by negation. This type of tweet is also included in the *NegCue* subset and in the *Total* set ($Total \supset NegCue \supset RuleAffect$).

> *Todo arranca de un tuit nada amable hacia Mou. Muchas de las críticas me descalificaban por ello como periodista en RNE. De ahí la pregunta.* (It all starts with an unkind tweet to Mou. Many of the criticisms blamed me for it as a journalist in RNE. Hence the question.)

TABLE 3
Results *Total* Set

	Macro-P	Macro-R	Macro-F1	Accuracy	Improvement Accuracy
BS	0.5764	0.5235	0.5486	0.6258	-
BSN	0.5705	0.5190	0.5435	0.6205	-0.885%
NR	0.5810	0.5296	0.5541	0.6308	0.80%

Note: The improvement in the Accuracy is measured over the BS method.

TABLE 4
Confusion Matrix BS Experiment with *Total* Set

	Predicted P	Predicted NEU	Predicted N	Recall
Real P	16,476	4,768	989	0.7410
Real NEU	511	446	348	0.3418
Real N	2,758	5,360	7,726	0.4876
Precision	0.8344	0.0422	0.8525	

6.2 Results

After these clarifications, the results achieved in the experiments with the *Total* set are shown in Table 3.

It can been seen that the integration of the most common approach to detect the scope of negation in English tweets (BSN), [48]) does not work well in the system that we use for the polarity classification of Spanish tweets. On the other hand, when the rules-based approach that we propose is included (NR), there is an improvement, but perhaps it seems that it is not so significant. However, if we observe the confusion matrix of the experiments (Table 4 and Table 5) we can see that there is a difference of about 200 tweets which have been correctly classified with the NR experiment.

As we have mentioned earlier, in order to evaluate the rules defined we should pay attention to the tweets with

TABLE 5
Confusion Matrix NR Experiment with *Total* Set

	Predicted P	Predicted NEU	Predicted N	Recall
Real P	16,566	4,746	921	0.7451
Real NEU	511	457	337	0.3502
Real N	2,685	5,341	7,818	0.4934
Precision	0.8383	0.0433	0.8614	

negative particles ($NegCue$) and mainly to the tweets with polar tokens in the scope of negation ($RuleAffected$). Tables 6 and 7 show the results obtained using these subsets.

As was to be expected, the values of the evaluation measures are lower than using the *Total* set (Table 3) because these subsets contain the most problematic tweets, i.e., the tweets with negation cues that are the most difficult to classify. In addition to these tweets, the *Total* set has other tweets without negation cues that are easier to classify, meaning that precision and recall increase. However, the improvement obtained with the rule-based approach is more evident. Furthermore, according to the results, it is reasserted the fact that the method most used to determine the scope of negation in English tweets (*BSN*) does not classify better in our system than the method that we propose for Spanish tweets (*NR*). The results achieved with the $RuleAffect$ subset show the evaluation of the rules that we have presented in this paper. Of course, the rules are not perfect and can be improved in order to increase accuracy. However, there is apparently a significant difference between *BS* and *NR* because as we can see there is an improvement of 18.57 percent in the accuracy and 19.19 percent in the F1 measure. Therefore, to avoid wrong conclusions we will perform a statistical analysis to check whether the rules defined for the treatment of negation really do improve the classification.

7 Analysis of Results

In order to see if there is a significant difference between the proportions of tweets correctly classified using the *BS* method and the *NR* method we have carried out a hypothesis test. There are different statistical tests that can be used to test the difference in the proportions of two populations depending on whether we are going to compare measurements that have been observed in separate (independent) groups or in the same group of subjects before and after an event (matched-pairs). The most commonly used tests for comparing two independent proportions are the Z-test and the Chi-square test. In the case of matched-pairs, the most frequently used tests are the Wilcoxon-signed rank test and the sign test for quantitative data, and McNemar's test for qualitative data. We have used McNemar's test because our

TABLE 6
Results *NegCue* Set

	Macro-P	Macro-R	Macro-F1	Accuracy	Improvement Accuracy
BS	0.4861	0.4702	0.4780	0.4866	-
BSN	0.4621	0.4514	0.4567	0.4622	-5.01%
NR	0.5060	0.4936	0.4997	0.5092	4.64%

Note: The improvement in the Accuracy is measured over the BS method.

TABLE 7
Results *RuleAffect* Set

	Macro-P	Macro-R	Macro-F1	Accuracy	Improvement Accuracy
BS	0.3971	0.3949	0.3960	0.4463	-
BSN	0.4431	0.4545	0.4487	0.5026	12.61%
NR	0.4660	0.4792	0.4725	0.5292	18.57%

Note: The improvement in the Accuracy is measured over the BS method.

data are qualitative and whenever possible it is better to work with data in the original form.

Formally, McNemar's test is known as a model for matched-pairs data with a binary response [49], [50]. This test is used to compare two proportions that have been observed in the same group of subjects, but at two different times (before and after an event), that is, it is used to determine if there are differences on a dichotomous dependent variable (i.e., correctly classified = {yes, no}) between two related groups (i.e., classifier = {BS method, NR method}). It attempts to compare whether there is any significant change between the two measurements.

We want to compare the proportions of opinions correctly classified before taking into account negation (BS method) and after considering it (NR method). Thus, we formulate the following one-sided hypothesis test, at the level of significance $\alpha = 0.01$, to verify if the proportion of tweets correctly classified using the NR method is greater than the proportion using the BS method

$$\begin{aligned} H_0 &: p_1 \leq p_2 \\ H_1 &: p_1 > p_2, \end{aligned} \tag{9}$$

where p_1 is the proportion of tweets correctly classified taking into account negation (*NR* model) and p_2 is the proportion of successes without treatment of negation (*BS* model).

We can assert at the level of significance $\alpha = 0.01$ that the proportions of tweets correctly classified taking into account negation with the rule-based method is significantly greater than the proportion of success using the classifier without considering negation (p-value = $2.6976 * 10^{-7} \approx 0$, McNemar's test, $\alpha = 0.01$).

It has been demonstrated that the method proposed to determine the scope of negation improves the polarity classification of Spanish tweets, but we also want to analyze how each of the rules has worked. *RuleAffect* is the subset in which the rules have been evaluated, so it is interesting to know the frequency of use of each of the negation cues studied (Table 8).

The most widely used negation cue is *no* at 2,517 times, which indicates that a proper treatment of it is very important. This cue appears in 2,086 tweets, approximately 90 percent of the tweets evaluated (Table 9). This accurately reflects that most of the sentences with negation in Spanish employ this cue.[11] Moreover, this also happens in English. The most common linguistic negation cue in English is *not*, along with contractions created with it, such as *couldn't* or *isn't* [52]. The classification results using only the rule for the negation cue *no* (NR_onlyno) are shown in Table 10. Although the improvement in performance is achieved

11. According to [51], "The most common grammatical negation procedure in Spanish is the use of the adverb *not* before the verb".

TABLE 8
Frequency of Negation Cues in *RuleAffect* Subset

Particle	Frequency
no (not)	2,517
tampoco (neither)	36
nadie (nobody)	54
jamás (never)	10
ni (nor)	248
sin (without)	195
nada (nothing)	114
nunca (never)	58
ninguno (none)	5

TABLE 9
Tweets per Negation Cue in the *RuleAffect* Subset

	Total
no (not)	2,086
tampoco (neither)	35
nadie (nobody)	54
jamás (never)	10
ni (nor)	181
sin (without)	182
nada (nothing)	107
nunca (never)	52
ninguno (none)	5

TABLE 10
Results *RuleAffect* Set-Comparative *NR_onlyno* and *NR* Approaches

	Macro-P	Macro-R	Macro-F1	Accuracy	Improvement Accuracy
NR_onlyno	0.4587	0.4693	0.4639	0.5155	-
NR	0.4660	0.4792	0.4725	0.5292	2.66%%

Note: NR_onlyno is the approach NR, but only applying the rule of the cue no.

TABLE 11
% Tweets per Negation Cue Correctly Classified (*RuleAffect* Subset)

	BS	NR
no (not)	44.10%	52.59%
tampoco (neither)	40.00%	54.29%
nadie (nobody)	40.74%	57.41%
jamás (never)	30.00%	20.00%
ni (nor)	35.36%	53.59%
sin (without)	51.65%	55.49%
nada (nothing)	34.58%	57.01%
nunca (never)	28.85%	63.46%
ninguno (none)	20.00%	20.00%

TABLE 12
Results *RuleAffect* Set-Comparative *NR* and *NR_mod* Approaches

	Macro-P	Macro-R	Macro-F1	Accuracy	Improvement Accuracy
NR	0.4660	0.4792	0.4725	0.5292	-
NR_Mod	0.4661	0.4797	0.4728	0.5297	0.09%

Note: NR_Mod is the approach NR, but without the application of the rules for the cues jamás *and* ninguno.

TABLE 13
Results *RuleAffect* Set-Comparative *NR* and *NR_Neu* Approaches

	Macro-P	Macro-R	Macro-F1	Accuracy	Improvement Accuracy
NR_Neu	0.4973	0.4404	0.4727	0.4304	-
NR	0.4660	0.4792	0.4725	0.5292	22.96%

Note: NR_Neu is the approach NR, but considering the polarity of the words that are in the scope of negation as neutral.

especially when the rule for the negation cue *no* is applied (which was to be expected because it is the most representative of the whole), when all the rules are considered the improvement is greater despite being small (which supports the validity of the proposed approach).

Table 11 shows the percentage of tweets per negation cue correctly classified with the approach that does not take into account negation (*BS*) and with the method that we propose to determine the scope of negation (*NR*). It can be seen that the rule-based method works better than the *BS* method for most of the cues, especially in the tweets with the cues *tampoco*, *nadie*, *ni*, *nada* and *nunca*, increasing the percentage of tweets correctly classified in a 34.61 percent in the best of the cases (cue *nunca*). Notwithstanding, it is noted that the rules for the cues *jamás* and *ninguno* do not work as well as we expected because the percentage of incorrectly classified tweets overcomes the percentage of correctly classified ones.

In order to know the reasons for the poor performance in the tweets with the negation cues *jamás* and *ninguno*, a deeper analysis was carried out. Most of the mistakes in the classification of these tweets have been produced in political tweets and in tweets belonging to the neutral class. We have to take into account that many political tweets are ironic and our system does not care about this challenge of SA. In order to see the influence of these two negation cues in the performance of the classification system, we carried out the experimentation without applying the rules defined in the tweets with the cues *jamás* and *ninguno* (NR_Mod). As can be noted the non-application of these rules improves the accuracy of the final system, although not by much (Table 12). This is due to the fact that the number of tweets with these cues represents a very small portion of the total. Therefore, it seems that the rules defined for these two negation cues are not suitable. Maybe, they need specific treatment. We will take this into account in our future research.

In relation to the other negation cues, the cases that lead to poor performance are due mainly to the presence of irony, double negation and the absence of certain polar words in the polarity lexicon used. Another reason could be the fact that when a word is in the scope of a negation cue our module always reverses its polarity, and the negation of a word does not always mean that its polarity is reversed, it can also be increased, decreased or considered as neutral. For example, "It is not perfect" is far from meaning "It is a disaster". Thus, although the results seem to confirm that the proposed heuristic is helpful in most cases, we conducted another experiment in which the polarity of the affected part was considered as neutral (Table 13). *NR_Neu*

TABLE 14
Confusion Matrix NR Experiment with *RuleAffect* Set

	Predicted P	Predicted NEU	Predicted N	Recall
Real P	472	115	99	0.6880
Real NEU	87	49	52	0.2606
Real N	459	283	710	0.4890
Precision	0.4637	0.1096	0.8246	

TABLE 15
Confusion Matrix NR_Neu Experiment with *RuleAffect* Set

	Predicted P	Predicted NEU	Predicted N	Recall
Real P	433	212	41	0.6312
Real NEU	79	71	38	0.3777
Real N	267	688	497	0.3423
Precision	0.5558	0.0731	0.8628	

heuristic increments Precision of positive and negative classes and Recall of neutral class, but Recall of positive and negative classes decreases because this approach misclassified some tweets as neutral (Tables 14 and 15). Consequently, the Accuracy of the approach that we propose in this study (*NR*) surpasses by 22.96 percent the Accuracy of the *NR_Neu* heuristic. However, it seems that it would be a good idea to study in which cases the polarity of the words that are within the scope of negation should be increased, decreased, swapped or considered as neutral rather than dealing with negation as a whole. Therefore, we are working on the annotation of negation and its scope [53] in the Spanish version of the SFU corpus [33] in order to study and discern these cases.

There is another important fact to discuss here and it is related to the chosen parser. The performance of traditional NLP tools is lower when tweets are analyzed, as is shown in several published papers [54], [55]. Research on English tweets is more advanced than on Spanish, so the fact that the first dependency parser for tweets written in English was presented in 2014 [54] is evidence that the task is not easy. Some studies describe the use of standard dependency parsers on tweets written in English [17] before the development of the specific dependency parser for tweets. As far as we know, there is no available dependency parser for Spanish tweets, but this cannot stop the advance of research. Although the use of specialized parsers is better for the processing of tweets, we also support the idea that while specialized parsers are not available, a standard parser can be used. In our study, the NLP tool most used for Spanish, Freeling, has been chosen.

8 Conclusions and Further Works

Negation is a linguistic phenomenon that can change the meaning of a sentence, so its treatment can influence positively in the performance of NLP tasks like SA. In this study, we have presented a set of syntactic rules for determining the scope of negation in Spanish. We have integrated these rules into a polarity classification system of Spanish tweets and it has been demonstrated that the results obtained with them are significantly greater than those without taking into account negation. This rule-based approach has also been compared with the method most used to determine the scope of negation in English tweets, and it has been proved that the classification with our approach is better. Moreover, we have analyzed the rules defined showing the performance of them with each negation cue.

The results obtained encourage us to follow in the study of the correct treatment of negation in the context of SA. However, one of the main problems in this area is the lack of resources. For example, there is no labeled corpus including negation for Spanish. Thus, we are currently working on the annotation of negation cues and their scope [53] in the Spanish version of the SFU corpus [33] in order to evaluate the rules with the aim of checking whether the system correctly determines the scope of the negation cues studied or if some of the errors are caused by the polarity classifier used. Moreover, we will also study in which cases the polarity of a word that is within the scope of negation should be swapped, considered neutral or if its value should be increased or decreased or, by contrast, whether it should not be changed.

Acknowledgments

This work has been partially supported by a Grant from the Ministerio de Educación, Cultura y Deporte (MECD-scholarship FPU014/00983), Fondo Europeo de Desarrollo Regional (FEDER) and REDES project (TIN2015-65136-C2-1-R) from the Ministerio de Economía y Competitividad.

References

[1] E. Martínez-Cámara, M. T. Martín-Valdivia, L. A. Ureña López, and A. Montejo-Raéz, "Sentiment analysis in Twitter," *Natural Language Eng.*, vol. 20, pp. 1–28, 1 2014.
[2] A. Go, R. Bhayani, and L. Huang, "Twitter sentiment classification using distant supervision," Stanford Univ., Stanford, CA, USA, Tech. Rep. CS224N, 2009. [Online]. Available: http://cs.stanford.edu/people/alecmgo/papers/TwitterDistantSupervision0 9.pdf
[3] F. Pla and L.-F. Hurtado, " Sentiment analysis in twitter for spanish," in *Natural Language Processing and Information Systems*, E. Métais, M. Roche, and M. Teisseire, Eds. Berlin, Germany: Springer, 2014, pp. 208–213.
[4] E. Martínez-Cámara, A. Montejo-Ráez, M. T. Martín-Valdivia, and L. A. Ureña-López, "SINAI: Machine learning and emotion of the crowd for sentiment analysis in microblogs," in *Proc. 2nd Joint Conf. Lexical Comput. Semantics, Volume 2: Proc. 7th Int. Workshop Semantic Eval.*, 2013, pp. 402–407. [Online]. Available: http://aclweb.org/anthology/S13–2066
[5] A. Montejo-Ráez, E. Martínez-Cámara, M. T. Martín-Valdivia, and L. A. Ureña López, "A knowledge-based approach for polarity classification in Twitter," *J. Assoc. Inf. Sci. Technol.*, vol. 65, no. 2, pp. 414–425, 2014. [Online]. Available: http://dx.doi.org/10.1002/asi.22984
[6] L. Zhang, R. Ghosh, M. Dekhil, M. Hsu, and B. Liu, "Combining lexicon-based and learning-based methods for twitter sentiment analysis," HP Lab., Palo Alto, CA, USA Tech. Rep. HPL-2011-89, Jun. 2011.
[7] A. Kumar and T. M. Sebastian, "Sentiment analysis on Twitter," *IJCSI Int. J. Comput. Sci. Issues*, vol. 9, no. 3, pp. 372–378, 2012.
[8] M. Thelwall, K. Buckley, and G. Paltoglou, "Sentiment strength detection for the social web," *J. American Soc. Inf. Sci. Technol.*, vol. 63, no. 1, pp. 163–173, 2012.
[9] S. Mohammad, S. Kiritchenko, and X. Zhu, "NRC-Canada: Building the state-of-the-art in sentiment analysis of tweets," in *Proc. 2nd Joint Conf. Lexical Comput. Semantics Volume 2: Proc. 7th Int. Workshop Semantic Eval.*, Jun. 2013, pp. 321–327. [Online]. Available: http://www.aclweb.org/anthology/S13–2053
[10] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up?: Sentiment classification using machine learning techniques," in *Proc. ACL-02 Conf. Empirical Methods Natural Language Process.*, 2002, pp. 79–86. [Online]. Available: http://dx.doi.org/10.3115/1118693.1118704

[11] A. Montejo-Ráez, “WeFeelFine as resource for unsupervised polarity classification,” *Procesamiento del Lenguaje Natural*, vol. 50, 2013, pp. 29–35. [Online]. Available: http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/4656

[12] S. D. Kamvar and J. Harris, “We feel fine and searching the emotional web,” in *Proc. 4th ACM Int. Conf. Web Search Data Mining*, 2011, pp. 117–126.

[13] A. Montejo-Ráez, M. C. Díaz-Galiano, and L. A. Ureña-López, “Crowd explicit sentiment analysis,” *Knowl.-Based Syst.*, vol. 69, pp. 134–139, 2014.

[14] D. Vilares, M. A. Alonso, and C. Gómez-Rodríguez, “On the usefulness of lexical and syntactic processing in polarity classification of Twitter messages,” *J. Assoc. Inf. Sci. Sci. Technol.*, vol. 66, pp. 1799–1816, 2015.

[15] J. Villena-Román, S. Lana-Serrano, E. Martínez-Cámara, and J. C. González-Cristóbal, “TASS-Workshop on sentiment analysis at SEPLN,” *Procesamiento del Lenguaje Natural*, vol. 50, 2013, pp. 37–44. [Online]. Available: http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/4657

[16] D. Maynard and A. Funk, “Automatic detection of political opinions in tweets,” in *The Semantic Web: ESWC 2011 Workshops*. Berlin, Germany: Springer, 2012, pp. 88–99.

[17] L. Jiang, M. Yu, M. Zhou, X. Liu, and T. Zhao, “Target-dependent Twitter sentiment classification,” in *Proc. 49th Annu. Meeting Assoc. Comput. Linguistics: Human Language Technol.*, vol. 1, 2011, pp. 151–160.

[18] V. Pérez-Rosas, C. Banea, and R. Mihalcea, “ Learning sentiment lexicons in spanish,” in *Proc. 8 Int. Conf. Language Resources Eval.*, May 2012, pp. 3077–3081.

[19] R. S. Trivedi and J. Eisenstein, “Discourse connectors for latent subjectivity in sentiment analysis,” in *Proc. Conf. North American Chapter Assoc. Comput. Linguistics: Human Language Technol.*, 2013, pp. 808–813.

[20] F. Batista and R. Ribeiro, “Sentiment analysis and topic classification based on binary maximum entropy classifiers,” *Procesamiento del Lenguaje Natural*, vol. 50, 2013, pp. 77–84. [Online]. Available: http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/4662

[21] G. Sidorov, et al., “Empirical study of machine learning based approach for opinion mining in tweets,” in *Proc. Advances Artif. Intell.*, vol. 7629, 2013, pp. 1–14.

[22] M. D. Molina-González, E. Martínez-Cámara, M. T. Martín-Valdivia, and J. M. Perea-Ortega, “Semantic orientation for polarity classification in Spanish reviews,” *Expert Syst. Appl.*, vol. 40, no. 18, pp. 7250–7257, Dec. 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0957417413004752

[23] M. D. Molina-González, E. Martínez-Cámara, M. T. Martín-Valdivia, and L. A. Ureña-López, “ Cross-domain sentiment analysis using Spanish opinionated words,” in *Proc. Natural Language Process. Inf. Syst.*, vol. 8455, 2014, pp. 214–219.

[24] E. Martínez-Cámara, M. T. Martín-Valdivia, M. D. Molina-González, and J. M. Perea-Ortega, “Integrating Spanish lexical resources by meta-classifiers for polarity classification,” *J. Inf. Sci.*, vol. 40, no. 4, pp. 538–554, 2014. [Online]. Available: http://jis.sagepub.com/content/40/4/538.abstract

[25] S. Das and M. Chen, “Yahoo! for Amazon: Extracting market sentiment from stock message boards,” in *Proc. Asia Pacific Finance Assoc. Annu. Conf.*, 2001, pp. 1–16.

[26] L. Polanyi and A. Zaenen, “Contextual lexical valence shifters,” in *Proc. AAAI Spring Symp. Exploring Attitude Affect Text: Theories Appl.*, 2004, pp. 1–10.

[27] A. Kennedy and D. Inkpen, “Sentiment classification of movie reviews using contextual valence shifters,” *Comput. Intell.*, vol. 22, no. 2, pp. 110–125, 2006. [Online]. Available: http://dx.doi.org/10.1111/j.1467–8640.2006.00277.x

[28] T. Wilson, J. Wiebe, and P. Hoffmann, “Recognizing contextual polarity in phrase-level sentiment analysis,” in *Proc. Conf. Human Language Technol. Empirical Methods Natural Language Process.*, 2005, pp. 347–354. [Online]. Available: http://dx.doi.org/10.3115/1220575.1220619

[29] L. Jia, C. Yu, and W. Meng, “The effect of negation on sentiment analysis and retrieval effectiveness,” in *Proc. 18th ACM Conf. Inf. Knowl. Manag.*, 2009, pp. 1827–1830. [Online]. Available: http://doi.acm.org/10.1145/1645953.1646241

[30] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede, “Lexicon-based methods for sentiment analysis,” *Comput. Linguistics*, vol. 37, no. 2, pp. 267–307, Jun. 2011. [Online]. Available: http://dx.doi.org/10.1162/COLI_a_00049

[31] R. Morante and C. Sporleder, “Modality and negation: An introduction to the special issue,” *Comput. Linguistics*, vol. 38, no. 2, pp. 223–260, Jun. 2012.

[32] M. Wiegand, A. Balahur, B. Roth, D. Klakow, and A. Montoyo, “A survey on the role of negation in sentiment analysis,” in *Proc. Workshop Negation Speculation Natural Language Process.*, 2010, pp. 60–68. [Online]. Available: http://dl.acm.org/citation.cfm?id=1858959.1858970

[33] J. Brooke, M. Tofiloski, and M. Taboada, “Cross-linguistic sentiment analysis: From English to Spanish,” in *Proc. Int. Conf.*, Sep. 2009, pp. 50–54. [Online]. Available: http://www.aclweb.org/anthology/R09–1010

[34] M. Taboada, K. Voll, and J. Brooke, “Extracting sentiment as a function of discourse structure and topicality,” School of Comput. Sci., Simon Fraser Univ., Burnaby, CA, Tech. Rep. 2008–20, Dec. 2, 2008.

[35] D. Vilares, M. Alonso, and C. Gmez-Rodrguez, “Polarity classification of opinionated Spanish texts using dependency parsing,” *Procesamiento del Lenguaje Natural*, vol. 50, no. 0, pp. 13–20, 2013. [Online]. Available: http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/4701

[36] E. Martínez-Cámara, M. T. Martín-Valdivia, L. A. Ureña López, and R. Mitkov, “Polarity classification for Spanish tweets using the COST corpus,” *J. Inf. Sci.*, vol. 41, pp. 263–272, 2015. [Online]. Available: http://jis.sagepub.com/content/early/2015/02/02/0165551514566564.abstract

[37] M. Hu and B. Liu, “Mining and summarizing customer reviews,” in *Proc. 10th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2004, pp. 168–177. [Online]. Available: http://doi.acm.org/10.1145/1014052.1014073

[38] P. J. Stone, D. C. Dunphy, and M. S. Smith, *The General Inquirer: A Computer Approach to Content Analysis*. Cambridge, MA, USA: MIT Press, 1966.

[39] S. Baccianella, A. Esuli, and F. Sebastiani, “ SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining,” in *Proc. 7th Int. Conf. Language Resources Eval.*, May 2010, pp. 2200–2204.

[40] S. M. Mohammad, “#Emotional Tweets,” in *Proc. 1st Joint Conf. Lexical Comput. Semantics - Volume 1: Proc. Main Conf. Shared Task Volume 2: Proc. 6th Int. Workshop Semantic Eval.*, 2012, pp. 246–255. [Online]. Available: http://dl.acm.org/citation.cfm?id=2387636.2387676

[41] J. Read, “Using emoticons to reduce dependency in machine learning techniques for sentiment classification,” in *Proc. ACL Student Res. Workshop*, 2005, pp. 43–48. [Online]. Available: http://portal.acm.org/citation.cfm?id=1628960.1628969

[42] R.A.E., *Nueva gramática de la lengua española. Morfología y sintaxis*. Madrid, Spain: Espasa Calpe, 2009.

[43] J. Atserias Batalla, E. Comelles Pujadas, and A. Mayor, “TXALA un analizador libre de dependencias para el castellano,” *Procesamiento del Lenguaje Natural*, vol. 35, no. 0, pp. 455–456, 2005. [Online]. Available: http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/3006

[44] L. Padró and E. Stanilovsky, “ FreeLing 3.0: Towards wider multilinguality,” in *Proc. 8th Int. Conf. Language Resources Eval.*, May 2012, pp. 2473–2479.

[45] T. Brants, “TnT: A statistical part-of-speech tagger,” in *Proc. 6th Conf. Appl. Natural Language Process.*, 2000, pp. 224–231.

[46] L. Padró, “A hybrid environment for syntax-semantic tagging,” *CoRR*, vol. cmp-lg/9802002, pp. 1–128, 1998. [Online]. Available: http://arxiv.org/abs/cmp-lg/9802002

[47] F. Sebastiani, “Machine learning in automated text categorization,” *ACM Comput. Surveys*, vol. 34, no. 1, pp. 1–47, Mar. 2002. [Online]. Available: http://doi.acm.org/10.1145/505282.505283

[48] C. Potts, “Sentiment symposium tutorial,” in *Sentiment Analysis Symposium*. San Francisco, CA, USA: Alta Plana Corporation, 2011. [Online]. Available: http://sentiment.christopherpotss.net/

[49] Q. McNemar, “ Note on the sampling error of the difference between correlated proportions or percentages,” *Psychometrika*, vol. 12, no. 2, pp. 153–157, 1947. [Online]. Available: http://dx.doi.org/10.1007/BF02295996

[50] A. Agresti, *An Introduction to Categorical Data Analysis*, 2nd ed., ser. Wiley Series in Probability and Statistics. Hoboken, NJ, USA: Wiley, Apr. 2007.

[51] M. G. Camarero, “Cuestiones pragmáticas sobre la negación,” *redELE: Revista Electrónica de Didáctica ELE*, vol. 5, no. 12, 2008, Art. no. 3.

[52] G. Tottie, "Negation in English speech and writing: A study in variation," *Language*, vol. 69, no. 3, pp. 590–593, 1993.
[53] M. A. Martí, M. T. Martín-Valdivia, M. Taulé, S. M. Jiménez-Zafra, M. Nofre, and L. Marsó, "La negación en español: Análisis y tipología de patrones de negación," *Procesamiento del Lenguaje Natural*, vol. 57, pp. 41–48, 2016.
[54] L. Kong, N. Schneider, S. Swayamdipta, A. Bhatia, C. Dyer, and N. A. Smith, "A dependency parser for tweets," in *Proc. Conf. Empirical Methods Natural Language Process.*, 2014, pp. 1001–1012.
[55] A. Stavrianou, C. Brun, T. Silander, and C. Roux, "NLP-based feature extraction for automated tweet classification," in *Proc. Interactions Data Mining Natural Language Process.*, 2014, Art. no. 145.

Salud María Jiménez-Zafra received the MS degree in computer science with the University of Jaén (Spain), in 2013, She is currently working toward the PhD degree in computer science at the University of Jaén (Spain). Her current research interest include related to Sentiment Analysis, especially with two of the challenges of this field, treatment of negation and aspect-level analysis.

M. Teresa Martín-Valdivia received the MS degree in computer science from the University of Granada, and the PhD degree in computer science from Software Engineering Department of Málaga. She is a Lecturer in the Department of Computer Science, University of Jaén (Spain). Her current research interests include Natural Language Processing, tools for text mining, neural networks, machine learning algorithms, text categorization, and multilingual and multimodal information retrieval. In addition, she is the Treasurer of SEPLN Committee since 2007 and editor of the Procesamiento de Lenguaje Natural Journal. She is author or co-author of more than 50 scientific publications. She is a technical reviewer for several journals and a member of the Program Committee of several major conferences.

Eugenio Martínez-Cámara received the MS and PhD degrees in computer science from the University of Jaén (Spain). His current research interests include related to several fields of natural language processing which include sentiment analysis, information extraction, semantic analysis, data analysis and machine learning.

L. Alfonso Ureña-López received the MS degree in computer science from the University of Granada and the PhD degree in computer science from the Software Engineering Department of Granada. He is professor in the Department of Computer Science, University of Jaén (Spain). His current main research interest include intelligent search, including search engines, search assistants, tools for text mining and retrieval and human computer Interaction. He is director and founder of the Research Group of Intelligent Systems on Information Access. He is also president of the Spanish NLP Society since 2007 and editor of the Procesamiento de Lenguaje Natural Journal. Currently, he is director of the Advanced Studies Center in Information and Communication Technologies at University of Jaén.