Evaluation of potential Spanish text markers on social posts as features for polarity classification

Edgar Casasola Murillo
Universidad de Costa Rica, Escuela de Ciencias de la Computación
San José, Costa Rica, 10501
edgar.casasola@ucr.ac.cr

Antonio Leoni de León
Universidad de Costa Rica, Departamento de Lingüística
San José, Costa Rica, 10501
antonio.leoni@ucr.ac.cr

and

Gabriela Marín Raventós
Universidad de Costa Rica, CITIC
San José, Costa Rica, 10501
gabriela.marin@ucr.ac.cr

Abstract

This work describes the identification and evaluation process of potential text markers for sentiment analysis. Evaluation of the markers and its use as part of the feature extraction process from plain text that is needed for sentiment analysis is presented. Evaluation of text marker obtained as a result of systematic analysis from a corpus over a second one allowed us to identify that emphasized positive words are strong indicators for positive text. The second corpus allowed us to evaluate the relation between the polarity of emphasized words and the text they appear in. Evaluation of the markers for polarity detection task in combination with a polarized dictionary produced polarity classification average precision of 56% using only three markers. This are promising results compared to the top 69% obtained using more features and specialized dictionaries for the same task.

Keywords: Information Retrieval, Natural Language Processing, Sentiment Analysis, Feature Vectors, Text Classification.

1 Introduction

The importance of Sentiment Analysis has become clear with the increasing amount of opinions posted online. Sentiment Analysis extracts information related to public opinion. It is a valuable resource to understand the way others perceive people's actions, services, products, institutions or events. Information is automatically processed and classified using algorithms such as SVM, Naive Bayes, Decision Trees among others. The main task related to Sentiment Analysis is the classification of text as positive or negative. This task is known as polarity detection.

According to [1] research has been moving from using classification based on single words to what is called feature extraction. The feature extraction techniques provide the basic input for polarity detection algorithms. The work of Chenlo [2] presents how the feature extraction techniques affect the results obtained by various classifiers in English. Recent research evaluates methods for feature extraction from text and its application for text classification [3], [4], [5], [6].

Normally, texts posted in Social Networks introduces variations that make its preprocessing difficult. Normalization of text needs to be done to correct grammatical errors and also to increase POS tagging performance [7]. But, sometimes important information aspects can get lost in the normalization process. Repetition of characters for exam-

ple denote exaggeration, but the terms are normally corrected to identify the word they refer to. The fact that repetition was present is not reported as a potential feature to be exploited.

This study is based on the quantification, evaluation and use of potential sentiment lexical markers for sentiment analysis. Potential markers are any superficial text form variations used to mark emphasis or other emotion related aspects as part of the text. The main goal was to identify and evaluate some potential markers that could be later be exploited at a feature extraction phase for polarity detection. This work is distinct from others because it illustrates how important it is to take into account a proper normalization of text, but also to preserve various text representations found to be important for Sentiment Analysis. This work includes a description of the systematic approach used to identify the sentiment markers, the individual evaluation of them when related to the polarity of text in a known sentiment annotated data set, and the finally an evaluation of its use for sentiment analysis [1].

There is an increased amount of research in Spanish Sentiment Analysis [8] [9]. By the year 2013 Spanish was the third most used language on the Internet `http://www.internetworldstats.com/stats7.htm`, but there is an existing research gap if compared against English Language as it is addressed in [10]. As an effort to narrow that gap our research focuses in Spanish. As a result, we used a Spanish Corpus of Facebook postings to identify potential text markers.

Traditional separation of the Natural Language Process into a series of steps was originated as a pedagogical aid and it became the basics for architectural models [11] used for language processing. This widely accepted approach consist of one stack of encapsulated and sequential stages. In this traditional process different levels of data abstraction are handled. The lower level correspond to finer-grained decomposition related to tokenization and sentence segmentation. At every stage, there is an expected product to be used as input at a higher stage. The disadvantage of such a modular approach is that on every transition some useful information can be disregarded.

The initial feature extraction techniques cited by Pang and Lee [12] were based on term presence, identification of sentiment words, bi-grams and syntax features related to the POS tagging function of words. Normal text preprocessing procedures such as the ones mentioned by [13], the forced lowercase of terms, the elimination of terms with low frequencies, and the elimination of stop words and stemming, became common practices for natural language processing applications.

From a classification point of view, if a word is taken by itself as a stand alone term then words that do not repeat as useless since they are not helpful for future classification purposes. Sometimes the initial normalization of text goes throw a process based on a computational perspective only as cited by Forman:

> "Easily half of the total number of distinct words may occur only a single time, so eliminating words under a given low rate of occurrence yields great savings" [13]

The question that rises here is: Why if this words appearing rarely have some morphological or lexical mark that could be summarized as a feature? Could the presence of some sentiment mark in the text be disregarded by assuming the way things have been done so far is the correct way to do it. What would happen if we identify those text markers at the lexical and syntactic stages, and use them later as part of the features for pragmatical purposes?

Architectural view of natural language processing as a series of steps preserving potential text markers. As shown in Fig. 1, normalization of text should preserve information about variations on the external form of text.

The next sections mention the procedure we use to identify potential markers. Later an evaluation of those markers were related to the polarity of text and the polarity of terms is presented. In that section, the use of a polarity dictionary and the TASS2015 training data `http://www.sngularmeaning.team/TASS2015/tass2015.php` set for the polarity detection task are presented as part of the evaluation of the potential text markers for sentiment analysis. Finally, some results about the evaluation of the markers to identify polarity is discussed.

2 Identification of the emergent variations of text in social media

Initially, a method based on a sequence of text processing steps with a dictionary centered approach was used to identify potential markers. An initial set of terms was created including all the terms appearing at the domain specific corpus to be analyzed. The classification of all terms was the final goal, categorizing them as well form dictionary words, emphasized or modified word, kind of emphasis or word-form variation used. Classification of terms was done by executing successive processing steps. Each step works using the remaining set of unclassified terms from the previous step. After a word is found in the dictionary it is excluded from the process and only the words not found in the dictionary are part of the next step set of words. A series of form manipulation operations are applied to a given term to convert it to a known dictionary term. This process is shown in Fig. 2.

A corpus made of 1,910,514 Spanish comments was used for potential marker quantification. Using this corpus potential lexical text markers are identified and count. The FCBKCR2013 Corpus was created extracting user posts from the most important sources of public opinion in Costa Rica according to [14]. Arce identified and ranked

[1]The paper is an extended version of the work presented at the First Colloquium of Natural Language Processing NLPCR2016 and selected for publication at CLEI Electronic Journal.

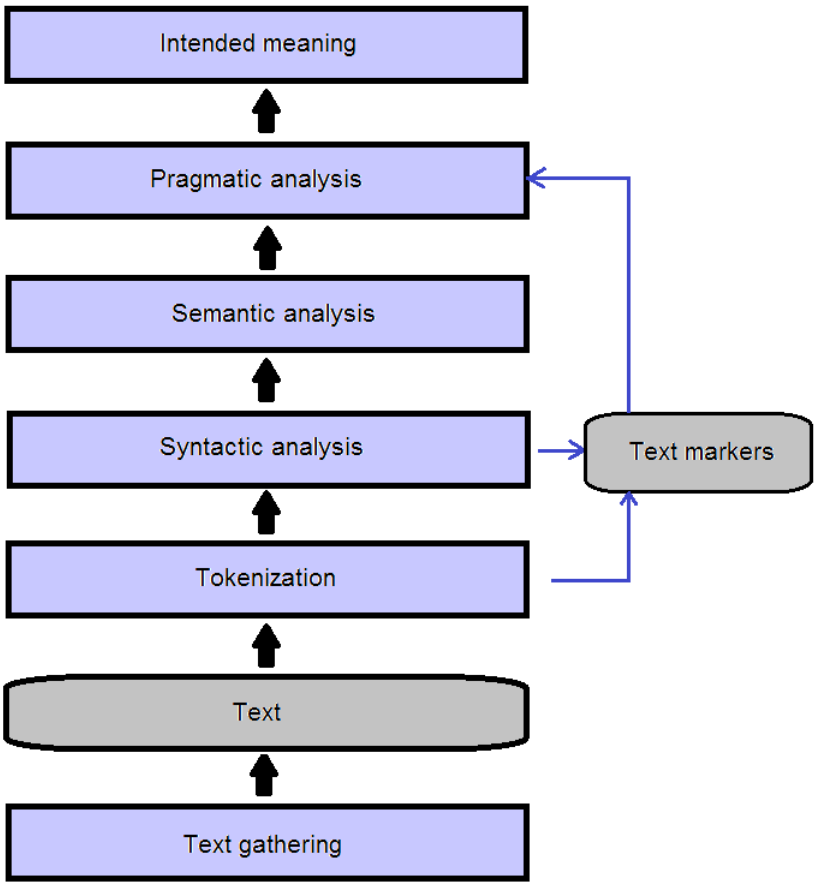

Figure 1: Extraction of potential markers can be done at the initial tokenization step

the most important sources of information about public matters in Costa Rica. These include Internet sites and the corresponding Facebook profiles owned by popular media TV, newspapers and popular independent radio programs in the country. The corpus was created collecting all posts in the year 2013 added to posts referring to any national event. All user comments posted as response to any publication at these twelve Facebook profiles were collected. The dictionary we use is the one available as part of the Open Office Project extension Spanish dictionary (http://extensions.openoffice.org/en/project/spanish-espanol). After applying the method to the corpus one set of words with any associated lexical mark was kept at each step. The first step of the process consisted on regular text parsing towards token identification. Tokens were searched in the dictionary and the ones appearing and not appearing in the dictionary were count. The absolute frequencies in the dictionary and in the corpus were count and their percentages estimated. Initially, 81.67% of the words in the Corpus appeared in the dictionary, and they represented 19.37% of the dictionary words. These words were identified as they appear in the text, without any lexical transformation. No uppercasing, accents or character repetition elimination was applied to the terms at this point. As expected approximately 20% of terms corresponded to 80% of the corpus. This Pareto relationship typically emerge when dealing with a significant amount of natural language data. Words not appearing in the dictionary were the modified words to be analyzed, they correspond to 18.33% of the words in the Corpus, as shown in Table 1.

Table 1: Percentage of text appearing in dictionary

Term in dicctionary	% of Terms	% of Corpus
Yes	19.37%	81.67%
No	80.63%	18.33%

As a regular preprocessing action, diacritic symbols were eliminated. This error correction related to missing accent symbols in the text allowed us to identify 5.17% of dictionary words. Those words, corresponded to 1.93% of words in the original Corpus leaving only 7.40% of Corpus terms for the analysis. Is here where the uppercasing markup, the repetition of characters and syllabic repetition was applied to the remaining words following the process previously described. During this phase we noticed that it is important to apply accent correction after the upper case character correction. Therefore, early elimination of accents could lead to an incorrect reduction of the term to the

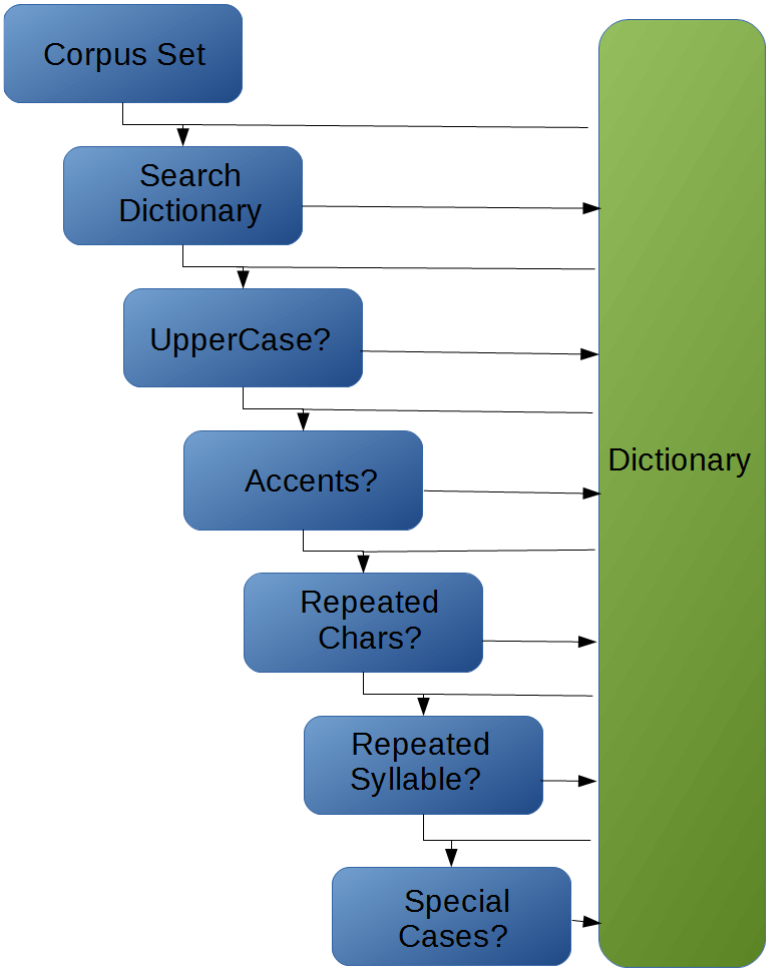

Figure 2: A dictionary based approach was used for term normalization and quantification of appearance of markers such as uppercasing, repetition of characters, syllabic repetition among others

correct word it represents. Moreover, at this step it was possible to observe that uppercase words on the corpus could be related to the user intention to emphasize its opinion.

Full uppercasing of words is considered a lexical mark denoting some kind of emphasis. The emphasis based on uppercasing is shown in the case of yelling or an exaggerated laugh or expression. Some frequent uppercase terms are shown in Table 2. Uppercase negation or expressions like YA could be related to some kind of emphasis possibly related to anger. The term muy used as an adverb amplifies the degree of expressions such as *muy feliz* (**very happy**) o *muy triste* (**very sad**). Its uppercase versions denote the intention to express an even higher intensity. The most frequent uppercase word in our Spanish corpus is the negation *NO* with a frequency of 32,567. Other words like *YA* or the onomatopoeia associated with the sound of laugh *JA* (**exaggerated laugh**) are frequently used.

Table 2: Most Frequent Uppercased Terms

Term	Frequency
NO (**NO**)	32567
YA (**NOW**)	3775
JA (**HA**)	2502
NUNCA (**NEVER**)	2327
MUY (**VERY**)	1513
ASCO (**DISGUST**)	991

Table 3 shows the percentage of words in the corpus using uppercase emphasis. Uppercased terms in the corpus correspond to 8.89% of the words it contains. Its coverage seems to be important if used to emphasize emotion.

Uppercased terms not appearing in the dictionary at this step normally require some extra normalization related to the elimination of duplicate character or syllabic repetition.

Table 3: Precentage of Uppercased Terms in the Corpus

Terms	Frequency	% of Uppercased Terms
59,348	3,484,917	8.89%

As Kouloumpis [7] mentions elimination of repeated characters when normalizing text is common. Those works eliminate repetitions, the issue here is that they do not mention exploiting such properties of text as part of the feature extraction process. Character or syllabic repetitions are indicators of exaggeration or vocal intonation. This repetition is also useful on exclamation **!** and interrogation signs **?**. For example de sequence **!!!** appeared 98,866 times and the sequence **???** appeared 27,512.

In Spanish, special cases are the characters: **c**, **r**, **l**, **n**, **e**, **o** and **z**. They can appear in pairs. The word version with repeated characters should be looked up first in the dictionary. If it is not recognized the version with removed repetition of characters should be searched for. Duplicity of these characters must not be removed automatically in proper names. The last name Murillo is an example. Some words frequently appearing with repeated characters are mention in Table 4.

Table 4: Frequent words with repeated characters

Terms	Frequency
no (**no**)	3,385
si (**yes**)	1,567
ah (**ah**)	1,544

Different variations of the word no are normally used. In this case Table 5 shows some variations and their frequency. Forms as noooo, Noooooooo or NOOOOO represent a different sentiment emphasis than a normal no. The frequencies or words with repeated characters in the studied corpus was of 26,499 and it was applied over 83,453 of the words in the dictionary , corresponding to 5.38% or the total of dictionary words. As revealed, variations of a term based on character repetition are not only frequent but are also sparse. Notice that as the number of the repeated character increases the frequency of that variation decreases. We can also observe at this point the mixing of uppercase emphasis and character repetition emphasis for the same word.

Table 5: Variations of the word with repeated characters**NO**

Term	Frequency
noooo	383
Noooo	351
nooooo	267
Nooooo	252
Noooooo	181
nooooooo	156
Noooooooo	131
NOOOOO	122
noooooooo	96

Other form of regulation of the meaning's intensity of words appears in the form of syllabic repetition. Use of a single *ja* (*ha!*) could be interpreted as an ironic laugh, *jaja* (*haha*) a sincere laugh, while repetitions such as *jajaja* (**hahaha**) are more like an intense laughing. The most frequent form of laughing in the corpus was the single *ja* (appearing 34,152 times) followed by *jajaja* 28,407 and *jajajaja* (15,223). Our normalization method detects the syllabic repetition and reduces the term to the basic form, but the mark recalling the repetition is kept as part of the sentiment markup of the dictionary term.

One interesting case to observe is the superlative *muchÃsimo* (**very** or **a lot**). The particularity of this word relates to its grammatical function as superlative. This word expresses the maximum degree on either its adjective or adverbial forms (**very**, **much** or **a lot**). It seems that the user perceives the superlative as not been strong enough or insufficient to express an specific characteristic to its right degree. The maximum seems not to be enough for the user. So, the user strategy is recurring to syllabic repetition to provide a very emphatic sense. For instance, words as: *muchÃsimo*, *muchÃsiiimas*, *muchisimaaaa*, *muchisimooooo* or *muchisisisisisima*; are just few of the many variation of the word *mucho*.

Consequently, the analysis of the corpus with the planned process allowed us to identify important sentiment text markers to be exploited. The following list mentions the ones we considered to be relevant for Spanish text from Facebook. These markers were implemented as part of our text normalizer:

1. Use of uppercase as word emphasis
2. Words modified by repetition of characters withing the word
3. Words modified with syllabic repetition
4. Symbol repetition as: ... , *!!!* or *???*
5. Emoticons
6. Identification of common abbreviations a acronyms.

Based on the previous process the normalization produces a clean version of text by expanding abbreviations, correcting accents and recognizing and correcting text with potential emphasis and sentiment related markets. The resulting text normalizer was implemented using the Java programming language. This implementation has the benefit of producing a clean version of the text comment it receives as input, and it optionally can produce the annotation of text for sentiment analysis.

A software API for normalization named *SpanishTextNormalizer* was written using the Java language. Among its functionality the Spanish text normalization modules has the following functions:

```
// Class Constructor
public SpanishTextNormalizer();

// Identification of emphasis
public boolean isUppercaseEmphasized(String term);
public boolean isCharRepEmphasized(String term);
public boolean isDupSylEmphasized(String term);

// String transformation functions for:

// Eliminate repetitions
public String charRepContraction(String term);
public String repSylContraction(String term);


// Some usefull domain specific functions
public String[] expandAbbreviation(String term);
public String normalizeWord(String term,boolean addTags);
public String normalizePhrase(String phrase, boolean addTags);
```

The software has the possibility to add tags, detect, or transform the texts it works with. The option named addTags is present as a parameter on some of the methods. This option activates tagging. Some extra functionality related to the normalization process and the handling of Spanish special characters and accents is algo available. The Spanish Text Normalizer has the option to validate dictionary words using Hunspell `https://hunspell.github.io/` as well as approximate match of words using an implementation of the popular Levenstein distance algorithm [11].

To illustrate the normalization process consider the following sample comment selected from the corpus:

> q decepcion y q mal hijo
> y asi como es con la mama
> asi va a ser con la doña
> Q MAAAAAAAAAAL

After normalization the resulting text is:

> qué decepción y qué mal hijo
> y así como es con la mamá
> así va a ser con la doña
> qué mal

Some of the resulting annotations using XML markup would be like this:

<ABREV> qué </ABREV> decepción y <ABREV> qué </ABREV> mal hijo.
y así como es con la mamá
así va a ser con la doña
qué <UPPER> <REP_CHAR>mal</REP_CHAR> </UPPER>

Notice that the annotation of the extended version of the abbreviations are mark, and the word *MAL* (**bad**) is marked as **UPPER** and **REPCHAR** because the presence of both, complete word uppercase, and repetition of characters. For evaluation purposes the first three markers shown in the previous list were selected. Subsequently, after identification of the potential sentiment markers an a proposal for their use became necessary. The next section describes the evaluation process for the individual markers.

3 Evaluation of individual markers

To evaluate the results obtained by exploiting some of this markers we used the TASS 2015 general collection used as part of the Polarity Classification Task[15]. This collection is made of 7217 polarity annotated tweets. We use 5068 posts with previously added polarity. Only positive (**P** and **P+**) and negative (**N** and **N+**) where taken into account. The 2885 positive tweets and 2182 negative tweets were process extracting words with each of the three emphasis to be evaluated. Therefore, the words using uppercase emphasis, repeated characters or syllabic repetition were extracted and store in separated files. Additionally, a polarity dictionary created with a variation of Turney method [16] was used.

The polarity dictionary was created using a variation of the **PMI-IR** version of the Turney algorithm as part of our research. As a variation from the original Turney method, we used 20 million tweets indexed using the open source search engine **SOLR** (http://lucene.apache.org/solr/) instead of using a Web search engine. On the other hand, the **NEAR** operation was omitted since the size of the tweets is small enough and normally contains one or at the most two sentences. The created dictionary is made of 15,174 dictionary terms with a real value polarity annotation. The polarity value is a real number between -1.0 and 1.0 ranging from the most negative -1.0 to the most positive 1.0.

Figure 3 shows the process used to extract information about every individual word. Words with emphasis from the TASS 2015 Data Set were normalized and extracted. Information about word polarity and the tweets they appear in and their annotated polarity was calculated. For each word, its emphasis, the dictionary polarity and the polarity of the tweets they appear in was registered. For each specific word the number of positive tweets and negative tweets they appear in was registered. The polarity of the word for each kind of emphasis was related to the percentage of tweets with the same polarity for the word and the average precision was calculated for the emphasis.

The results for each emphasis is shown in Tables 6,7 and 8. The idea of this analysis was to detect the possible correlation among positive emphasized words with positive tweets and the one of negative emphasized words with negative tweets.

Table 6: Proportion of tweets with character repetition words

	Positive tweets	Negative tweets
Char repetition positive terms	1.0	0
Char repetition negative terms	0.5652173913	0.435

Observing the relation between positive an negative emphasized words with positive and negative tweets an interesting characteristic appears. We can observe that for the three emphasis, when a positive word is emphasized there is a high chance that the tweet is positive. This correlation doesn't appear to maintain for the negative emphasized words appearing in negative tweets.

Table 7: Proportion of tweets with syllabic repetition words

	Positive class	Negative class
Syl. repetition positive terms	0.818	0.18181818
Syl. repetition negative terms	0.489	0.4107142857

Positive words using emphasized repetition of characters according with our general purpose dictionary tend to appear on positive tweets. This results doesn't hold in the case of negative terms. Furthermore, when observing results for the other kind of emphasis we obtain a precision of 0.818 for syllabic repetition emphasized terms and 0.804 for uppercased positive words. The average precision over positive tweets using positive dictionary terms as an indication for the TASS data set was of 0.874.

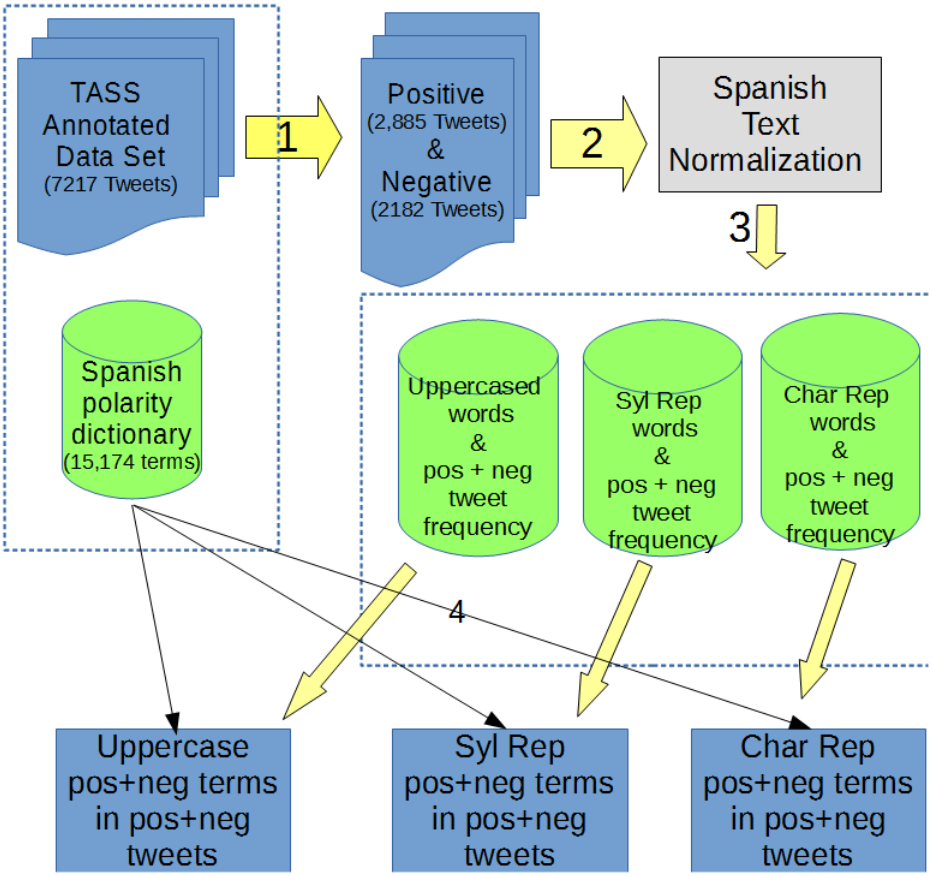

Figure 3: Evaluation of emphasized words

Table 8: Proportion of tweets with uppercased words

	Positive class	Negative class
Uppercased positive terms	0.8048780488	0.1951219512
Uppercased negative terms	0.4514285714	0.5485714286

Other important direct observation obtained from the data is the differentiated behavior of positive and negative opinions. Considering positive and negative opinions as opposite an symmetric manifestations has been one important misleading assumption in sentiment analysis. Classification of positive and negative opinions should be analyzed on their own.

An important potential use of this results related to the use of such emphasized words as seed for the creation of domain specific polarized dictionaries. Mechanisms as the Turney methods used for the creation of dictionaries or bootstrapping techniques such as the ones mention in [9] could take into account emphasized words because of their potential correlation with positive texts.

Some of the most frequent emphasized positive term found on positive tweets were:

- enhorabuena
- besos
- felicidades
- encanta
- beso
- regalo
- deseo
- historias
- homenaje
- gira
- amor
- genial
- recomendable
- maravilla
- abrazos

Based on the previous results it was possible to observe that positive term with emphasis are potential markers for sentiment analysis. Nevertheless, the sparsity of features is known as been one of the main problems for machine learning techniques, for that reason the combined use of a dictionary and the features is proposed here. The use of the polarity dictionary combined with the markers could influence the precision for the polarity classification task.

In the next section the results of one experiment evaluating only the use of the polarity dictionary and the uppercase, char repetition, and syllabic repetition marks for text classification are presented. The results obtained after training a support vector machine classifier with vectors of features using the emphasis markers is presented.

The purpose of that final evaluation is to show how to combine the use of the emphasis marks as features and how to combine them with a dictionary to create a vector representation for document classification. Of course, the idea here is not to outperform the results obtained by other more complex techniques but to reveal the potential of detecting and using this specific features.

4 Evaluation of markers for polarity classification

To evaluate the markers feature vectors were created to represent the tweets to be classified. As before, only positive and negative tweets were selected for the evaluation. We assume a assemble classifier is to be used in order to concentrate on polarity classification.

In order not to use all words as features a reduced dimension vector was created for each tweet. For each evaluated emphasis, a vector of size 20 was generated for each post. The vector used for classification was obtained by the concatenation of the three vectors. Every variable represented the accumulated frequency distributed by polarity from -1 o 1 on steps of 0.1. Similar vectors were created for the features analyzed. Only character repetition, syllabic repetition and uppercase markers were taken into account.

The representation or vectorization process is shown in Fig. 4. The vector entries were distributed according to the polarity of the term emphasized. All vector have the same size for every tweet. The position depends on the polarity of the word and the emphasis. When two emphasis are present for a single word it appears in both positions. Although evaluation of other representation models is necessary, those are beyond the scope of this work.

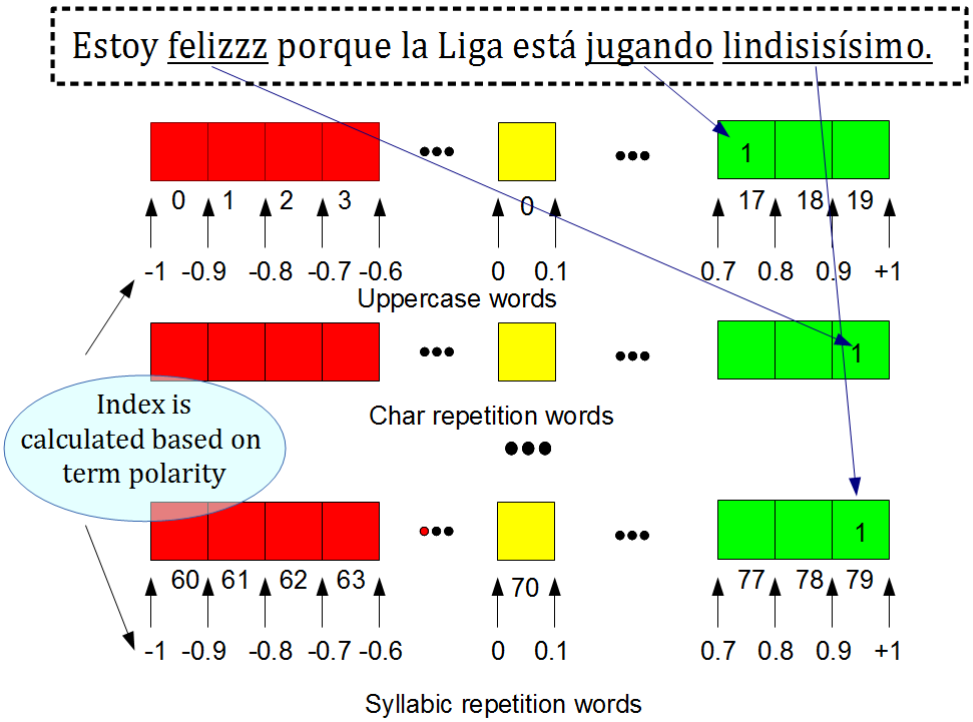

Figure 4: A feature vector was created for each tweet

The results were run using a SVM classifier libSVM as part of WEKA GNU open source software available at `http://www.cs.waikato.ac.nz/ml/weka/`. The trained model obtained by the previous training, was submited to the TASS competition where the model was evaluated against other systems. When training the classifier a 10 fold cross valitadion was used. On the training set the number of correctly classified instances was 3567 corresponding to 70.3828 % of the instances. The number of incorrectly classified instances was 1501, it corresponds to 29.6172 % of the evaluated tweets. Precision for the positive class reached a 71% and 62% for the negative class. Recall on the other hand was 0.8 for the positive class and 0.7 for the negative class. The average precision and recall were both 0.7 (see Table 9). Those results only used a polarity dictionary and the text marked features as mention before. This model was then used to evaluate the test sets submitted to the TASS 2015 workshop.

Table 9: Polarity classification precision and recall

	Positive class	Negative class
Precision	0.71	0.62
Recall	0.8	0.7

The results at TASS were not as high as with the training set. With the 1K data set for the 3 category task the

average precision of the classifier was 56%. That value can be compared to the top average of 69% obtained by the best team at TASS for the same task. This was a promising result since our model only made use the three explored characteristics (the emphasis markers and the polarity dictionary) while the approaches used by other teams were a combination of more computationally complex models and used more resources.

5 Conclusions and future work

Application of a systematic analysis of a complete Spanish corpus of user postings from social media allowed us to identify and quantify the frequency of morphological text markers present in the tokenized terms. We found that special text characteristics in our corpus had a potential use as features for sentiment analysis. We propose to keep this information at early preprocessing stages to avoid loosing them if a traditional preprocessing is applied is important.

Evidence shows that using positive emphasized words as features is a promising technique worth to include as part of a regular features extraction process for the polarity classification of text. The combined use of dictionary polarity and emphasis to create feature vectors was illustrated.

The replication of this study to other languages could lead to evaluate the pertinence of this tags on their own languages. We consider that domain specific and language specific issues require analysis of corpus text before implementing the process normalization and feature extraction modules in real world sentiment analysis applications. In this case real user text from the social network was analyzed and nevertheless gave relative good results when evaluating it over a data set made of tweets. An average precision of 56% was reached at the polarity detection task using the TASS 2015 data sets.

This results shows that the identified markers are useful for the polarity detection sentiment analysis tasks. Meanwhile, more effort has to be done to better exploit these features at the pragmatic level, exploring new vectorization techniques and classification approaches. Also the correlation between positive emphasized terms to identify positive tweets should be study using other data sets.

As future work we expect to improve results obtained at polarity classification task with new text features based on this markers and new polarity recalculation mechanisms based on their polarity and their distribution over text. Also, a method for the creation of domain specific polarized dictionaries using automatic gathering of emphasized seed words could lead to promising results.

Acknowledgment

This work has been supported in part by the Universidad de Costa Rica (UCR) and the Ministerio de Ciencia, Tecnología y Telecomunicaciones (MICITT) de la República de Costa Rica.

References

[1] E. Cambria, R. Speer, C. Havasi, and A. Hussain, "Senticnet: A publicly available semantic resource for opinion mining." in *AAAI fall symposium: commonsense knowledge*, vol. 10, 2010, p. 02.

[2] J. M. Chenlo and D. E. Losada, "An empirical study of sentence features for subjectivity and polarity classification," *Information Sciences*, vol. 280, pp. 275–288, 2014. [Online]. Available: https://doi.org/10.1016/j.ins.2014.05.009

[3] M. A. Cabanlit and K. Junshean Espinosa, "Optimizing n-gram based text feature selection in sentiment analysis for commercial products in twitter through polarity lexicons," in *Information, Intelligence, Systems and Applications, IISA 2014, The 5th International Conference on*. IEEE, 2014, pp. 94–97. [Online]. Available: https://doi.org/10.1109/iisa.2014.6878767

[4] R. Feldman, "Techniques and applications for sentiment analysis," *Commun. ACM*, vol. 56, no. 4, pp. 82–89, Apr. 2013. [Online]. Available: http://doi.acm.org/10.1145/2436256.2436274

[5] L. Guo and X. Wan, "Exploiting syntactic and semantic relationships between terms for opinion retrieval," *Journal of the american society for information science and technology*, vol. 63, no. 11, pp. 2269–2282, Nov. 2012. [Online]. Available: https://doi.org/10.1002/asi.22724

[6] A. Sharma and S. Dey, "A comparative study of feature selection and machine learning techniques for sentiment analysis," in *Proceedings of the 2012 ACM Research in Applied Computation Symposium*. ACM, 2012, pp. 1–7. [Online]. Available: https://doi.org/10.1145/2401603.2401605

[7] E. Kouloumpis, T. Wilson, and J. D. Moore, "Twitter sentiment analysis: The good the bad and the omg!" *Icwsm*, vol. 11, pp. 538–541, 2011.

[8] M.-T. Martín-Valdivia, E. Martínez-Cámara, J.-M. Perea-Ortega, and L. A. Ureña-López, "Sentiment polarity detection in Spanish reviews combining supervised and unsupervised approaches," *Expert Systems with Applications*, vol. 40, no. 10, pp. 3934 – 3942, 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0957417412013267

[9] V. Pérez-Rosas, C. Banea, and R. Mihalcea, "Learning sentiment lexicons in spanish." in *LREC*, 2012, pp. 3077–3081.

[10] M. Melero, A. B. i Cardús, A. Moreno, G. Rehm, K. de Smedt, and H. Uszkoreit, *The Spanish language in the digital age*. Springer, 2012.

[11] N. Indurkhya and F. J. Damerau, *Handbook of natural language processing*. CRC Press, 2010, vol. 2.

[12] B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Foundations and trends in information retrieval*, vol. 2, no. 1-2, pp. 1–135, 2008. [Online]. Available: https://doi.org/10.1561/1500000011

[13] G. Forman, "An extensive empirical study of feature selection metrics for text classification," *The Journal of machine learning research*, vol. 3, pp. 1289–1305, 2003.

[14] J. L. Arce, "Medios de comunicación de masas en Costa Rica: Entre la digitalización, la convergencia y el auge de los "new media"," in *Hacia la Sociedad de la Información y el Conocimiento*. Programa Sociedad de la Información y el Conocimiento, Universidad de Costa Rica, 2012, pp. 283–308.

[15] J. Villena-Román, J. García Morera, M. García-Cumbreras, Á.-E. Martínez-Cámara, T. Martín-Valdivia, and A. Ureña-López, "Overview of tass 2015," in *Proceedings del Taller TASS 2015 en Análisis de Sentimiento de la XXXI Conferencia SEPLN 2015*, 2015, pp. 13–21.

[16] P. D. Turney, "Expressing implicit semantic relations without supervision," in *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, ser. ACL-44, no. July. Stroudsburg, PA, USA: Association for Computational Linguistics, 2006, pp. 313–320, published in: Proceeding ACL-44 Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics Pages 313-320 Association for Computational Linguistics Stroudsburg, PA, USA Â©2006. [Online]. Available: http://dx.doi.org/10.3115/1220175.1220215