Káñina, Rev. Artes y Letras, Univ. Costa Rica XL (Extraordinario): 21-32, 2016 / ISSN:2215-2636

EVALUATION OF POTENTIAL FEATURES PRESENT IN SHORT TEXTS IN SPANISH IN ORDER TO CLASSIFY THEM BY POLARITY

EVALUACIÓN DE CARACTERÍSTICAS POTENCIALES PRESENTES EN TEXTOS CORTOS EN ESPAÑOL PARA CLASIFICARLOS POR POLARIDAD

Édgar Casasola Murillo[1]

Antonio Leoni de León[2]

Gabriela Marín Raventós[3]

ABSTRACT

This work describes the identification and evaluation process of potential text markers for sentiment analysis. The evaluation of the markers and their use as part of the feature extraction process from plain text that is needed for sentiment analysis is presented. The evaluation of text markers obtained as a result of systematic analysis from a corpus over a second one allowed us to identify that emphasized positive words that tend to appear in positive text posts. The second corpus allowed us to evaluate the relation between the polarity of morphological text markers and the text they appear in. The evaluation of the markers for polarity detection task, in combination with a polarized dictionary, produced polarity classification average precision of 0.56 % using only three markers. These are promising results if we compared them to the top 0.69 % obtained using more features and specialized dictionaries for the same task.
Key Words: sentiment analysis, information gain, feature vectors, polarity, classification.

RESUMEN

Este trabajo describe el proceso de identificación y evaluación de marcadores potenciales de texto para análisis de sentimientos. Se presenta la evaluación de los marcadores y se propone la forma de utilizarlos para análisis de sentimientos. La evaluación de los marcadores identificados como producto del análisis sistemático de un primer corpus sobre otro nos permitió determinar que palabras positivas con énfasis tienden a aparecer principalmente en comentarios positivos. Con el segundo corpus, se evaluó la relación entre la polaridad de las

1 Universidad de Costa Rica. Escuela de Ciencias de la Computación, Programa de Posgrado en Computación e Informática y Centro de Investigaciones en Tecnologías de la Información y Comunicación (CITIC). Costa Rica. Correo electrónico: edgar.casasola@ucr.ac.cr

2 Universidad de Costa Rica, Profesor, Escuela de Filología, Lingüística y Literatura. Costa Rica. Correo electrónico: antonio.leoni@uc.ac.cr

3 Universidad de Costa Rica. Centro de Investigaciones en Tecnologías de la Información y Comunicación (CITIC). Costa Rica. Correo electrónico: gabrielamarinraventos@gmail.com
Recepción: 15/1/2016 Aceptación: 16/3/2016

palabras con énfasis y sus textos. Finalmente, se llevó a cabo una evaluación del uso de los marcadores sobre la tarea de identificación de polaridad de textos, con lo cual se obtuvo una precisión de 0.56 usando solo tres marcadores y un diccionario polarizado. Los resultados fueron prometedores en comparación con 0.69 % que fue la precisión más alta obtenida en la misma tarea mediante el uso de mayor cantidad de características y diccionarios especializados.
Palabras clave: análisis de sentimientos, ganancia de información, vectores de características, clasificación, polaridad.

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 form others perceive people's actions, services, products, institutions or events. Information is automatically processed and classified using algorithms such as SVM (support vector machines), 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 (Cambria *et al*., 2013), 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 (2014) 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 (Cabanlit and Espinosa, 2014; Feldman, 2013; Guo and Wan, 2012; Sharma and Dey, 2012).

Normally, texts posted in Social Networks introduce variations that make their preprocessing difficult. Normalization of text needs to be done to correct grammatical errors and also to increase POS tagging (part-of-speech tagging) performance (Kouloumpis, 2011). However, sometimes important information aspects can get lost in the normalization process. Repetition of characters, for example, are normally corrected to identify the word they refer to. The fact that repetition was present is not reported as part of the features extraction process.

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 later be exploited at a feature extraction phase for polarity detection. This work is distinctive 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 is found to be important for Sentiment Analysis. A description of the systematic approach is used to identify the sentiment markers, their individual evaluation, related to the polarity of a single text in a sentiment annotated data set is given, and finally, there is an evaluation of their use for sentiment analysis in the following pages.

Given the increased amount of research in Spanish Sentiment Analysis (Martín-Valdivia *et al.,* 2013; Perez-Rosas, 2012) we used a Spanish corpus of Facebook postings to identify potential text markers. This is important since by 2013 Spanish was the third most used language on the Internet (Stats, 2013), but there is an existing research gap if compared against English language as it is addressed in Melero *et al.* (2012).

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 (Indurkhya, 2010) used for language processing. This widely accepted approach consists of one stack of encapsulated and sequential stages. In this traditional process, different levels of data abstraction are handled. The lower level corresponds 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 (2008) 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 Forman (2003), 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 standalone term, then words that do not repeat are 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 (2003: 1291): "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." The questions that rise here are: What if these words rarely appearing 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 pragmatic purposes?

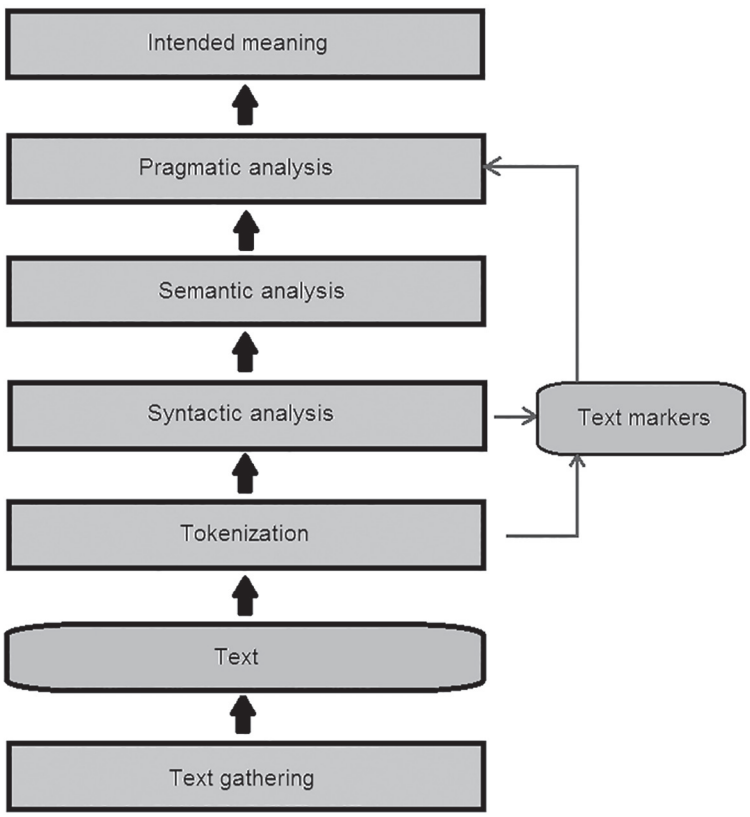

FIGURE 1

Extraction of potential markers can be done at the initial tokenization step. Normalization of text should preserve information about variations on the external form of text.

The next sections will mention the identification procedure we used to identify potential markers. Later, it is presented an evaluation of those markers, when related to the polarity of text and the polarity of terms. In that section, the use of a polarity dictionary and a known data set for the polarity detection task will be explained, 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 will be presented.

2. Identification of the emergent variations of text in social media

First, 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-formed dictionary words, emphasized or modified words, kind of emphasis or word-form variation used. The 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.

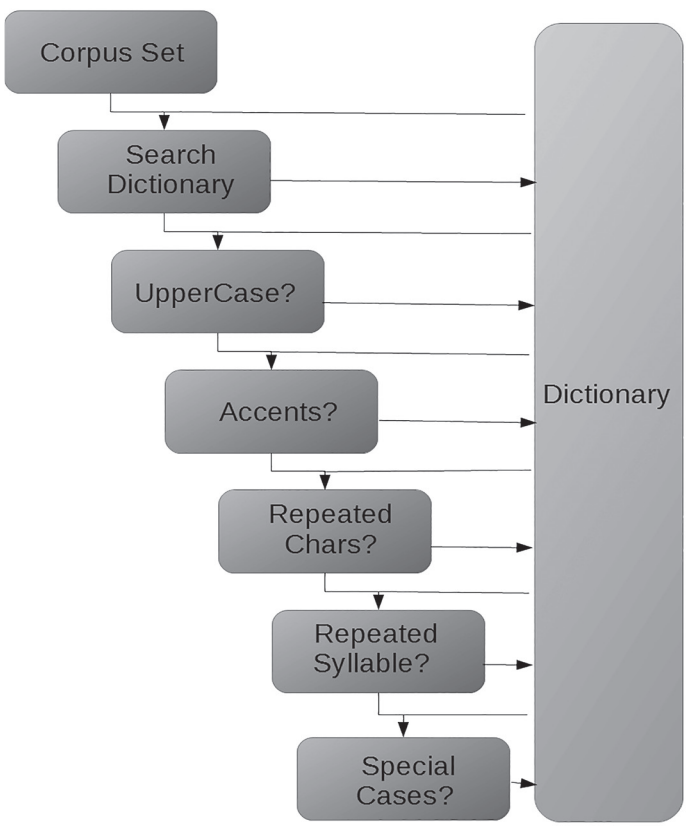

FIGURE 2

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.

A corpus made of **1,910,514** Spanish comments was used for potential marker quantification. In using this corpus, potential lexical text markers are identified and counted. The **FCBKCR2013 corpus** was created extracting user posts from the most important sources of public opinion in Costa Rica, according to Arce (2012). This author identified and ranked 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 added to posts referring to any national event during 2013. All user comments posted as response to any publication at these twelve Facebook profiles were collected. The dictionary we used is the one available as part of the Open Office Project extension Spanish dictionary[1].

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, both appearing and not appearing in the dictionary, were counted. The absolute frequencies in the dictionary and in the corpus were counted 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 emerges 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.

TABLE 1

Percentage of text appearing in the dictionary

Term in dictionary	% 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. It 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 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's intention to emphasize his/her 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 Upper Case Terms

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

Table 2 shows the frequency of upper cased words in the corpus. Nearly 8.89 % words in the corpus are versions of uppercase dictionary words. This appears to be an important marker denoting emotion. Terms that remain uppercased without appearing in the dictionary normally require some kind of elimination of duplicate or sequential character repetition. Other uppercase words not found in the dictionary needed some syllable repetition elimination or a combined elimination of repeated characters and repeated syllables.

TABLE 3

Percentage of uppercase terms in the corpus

Terms	Frequency	% of Terms
59,348	3,484,917	12.06 %

As Kouloumpis (2011) mentions, the 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, the characters {c, r, l, n, e, o and z} are special cases. They can appear in pairs. The word version of the word using two characters should be looked up first in the dictionary. If it is not recognized, the version with removed repetition of characters should be tried. Duplicity of these characters must not be removed automatically in proper names. The last name *Murillo* is an example.

Table 4 shows the frequency of terms with repeated characters.

TABLE 4

Percentage of words with repeated characters

Terms	Frequency
no (no)	3,385
sí (yes)	1,567
ah (ah)	1,544
muchísimo (too much)	18

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 were of 26,499 and it was applied to over 83,453 of the words in the dictionary, corresponding to 5.38 % of the total of dictionary words. As revealed, variations of a term based on character repetition are not only frequent but are also sparse. It is interesting to observe that as the number of the repeated character increases, the frequency of that variation decreases. We can also observe mixing of uppercase emphasis and character repetition emphasis for the same word.

TABLE 5

Frequency of NO with repeated character emphasis

Terms	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. The use of a single 'ja' (*ha!*) could be interpreted as an ironic laugh; 'jaja' (*haha*) a sincere laugh; while repetitions such as 'jajaja**'** (*hahaha*) is more like to represent 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 is the superlative 'muchísimo**'** (very much or a lot). The particularity of this word is because the grammatical function of the superlative is to express the maximum, on either its adjective or adverbial forms (very, much or a lot). It seems that the user perceives the superlative as insufficient to express specific characteristic on its right degree. The maximum seems to be not enough for the user. So, the user's strategy is recurring to syllabic repetition to provide a very emphatic sense. For instance, words as: 'muchísimo, muchisiiimas, muchisiiimos, muchisiimas, muchisimaaaa, muchisimoooo, muchisimooooo, muchisimoss, muchisisisima, muchisisisisisima' are just few of the many variation of the word *mucho*. Analyzing 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 were the ones implemented as part of our text normalizer:

1. Use of uppercase as word emphasis.
2. Words modified by repetition of characters within the word.
3. Words modified with syllabic repetition.
4. Symbol repetition as: ..., !!! or ???
5. Emoticons.

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 produces the annotation of text for sentiment analysis.

To illustrate the process, consider the following sample comment found in the corpus:

(a)
q decepcion y q mal hijo
y asi como es con la mama
asi va a ser con la doña
Q MAAAAAAAAAAAL.

After normalization the resulting text is:

(b)
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:

(c)
<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 marked, and the word 'mal' (*bad*) is marked as **UPPER** and **REPCHAR** because of 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, 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 these markers, we used the TASS 2014[2] general collection as part of the polarity classification task. This collection is made of 7217 polarity annotated Tweets instead of Facebook posts. We use 5068 posts with previously added polarity. Only positive (P and P+) and negative (N and N+) were taken into account. The 2885 positive Tweets and 2182 negative Tweets were processed extracting words with each of the three emphases 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's method was used (Turney, 2002). This dictionary was created using a variation of the PMI-IR version of the Turney algorithm as part of our research; it used 20 million tweets indexed using the open source search engine SOLR[3] instead of the Web, and the NEAR operation was omitted, since the size of the tweets is small enough and it normally contains one or at the most two sentences. It is made of 15,174 dictionary terms with a real value polarity annotation. The polarity value is a real number between -1.0, being the most negative, and 1.0 the most positive.

Figure 3 shows the process used to extract information about every individual word. For each word, its emphasis, the dictionary polarity, and the polarity of the tweets they appeared in were registered. For each specific word, the number of positive and negative tweets they were present in was calculated. 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.

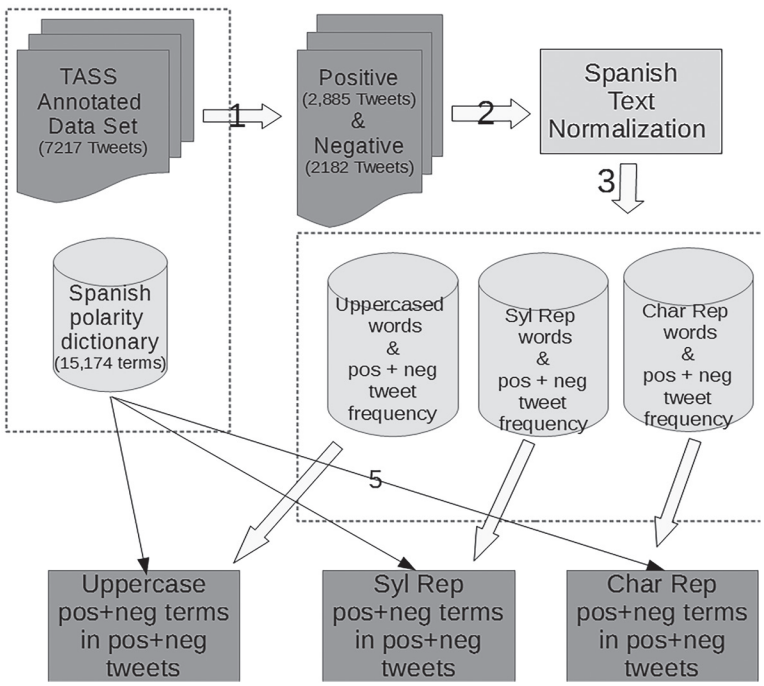

FIGURE 3

Evaluation of emphasized words. 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.

The results for each emphasis are shown in tables 6, 7, and 8 of the results. The most interesting observation is that, for the three emphases, 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 6

Proportion of Tweets polarity containing positive and negative emphasized words using character repetition

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

TABLE 7

Proportion of Tweets polarity containing positive and negative emphasized words using syllabic repetition

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

TABLE 8

Proportion of Tweets polarity containing positive and negative emphasized words using uppercasing

	Positive class	Negative class
REP. UPPER (POS. TERM)	0.8048780488	0.1951219512
REP. UPPER (NEG. TERM)	0.4514285714	0.5485714286

Based on the previous results, it was possible to observe that positive terms with emphasis are potential markers for sentiment analysis. 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 shown.

4. Results

To evaluate the markers, feature vectors were created to represent the tweets that are to be classified. As before, only positive and negative tweets were selected for the evaluation. In order to not 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 or 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. This vectorization process is shown in figure 4.

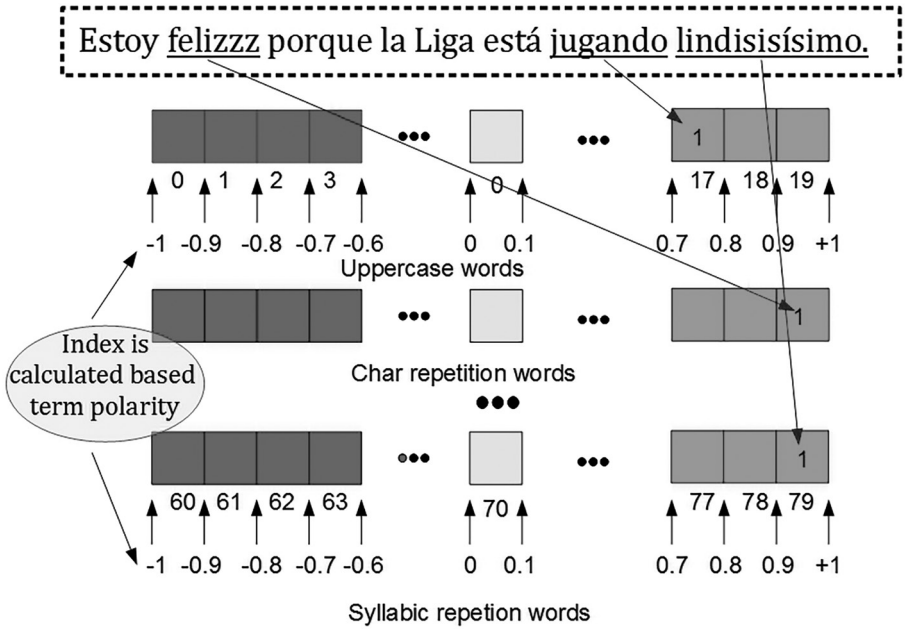

FIGURE 4

Vectorization. A feature vector was created for each tweet. The vector entries were distributed according to the polarity of the term emphasized. All vectors have the same size for every tweet. The position depends on the polarity of the word.

The results were run using a SVM classifier libSVM as part of WEKA GNU open source software[4]. Finally, the model obtained by the previous training was submitted to the TASS competition where the model was evaluated against other systems. When training the classifier, a 10 fold cross validation 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. The precision for the positive class reached a 0.71 % and 0.62 % for the negative class. The 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. Those results only used a polarity dictionary with the text marked features. This model was then used to evaluate the test sets submitted to the TASS 2015 workshop[5].

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 0.56 %. That value can be compared to the top average of 0.69 % obtained by the best team for the same task. This was a promising result since our model only used the three explored characteristics (the emphasis markers and the polarity dictionary), while other models used more complex and many different feature representations and classification implementations.

5. Conclusions and Future Work

The application of a systematic analysis of a complete Spanish corpus of Facebook postings 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 losing them if a traditional preprocessing is applied.

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

The replication of this study to other languages could lead to evaluate the pertinence of these tags on their own languages. We consider that domain specific and language specific issues require analysis of corpus text before implementing the process of normalization and feature extraction modules in real world sentiment analysis applications. In this case, real user text from the social network Facebook was analyzed and, even so, gave relative good results when evaluating it over a data set made of tweets. An average precision of 0.56 % was reached at the polarity detection task using the TASS 2015 data sets.

These results show 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 studied using other data sets.

As future work we expect to improve results obtained at polarity classification task with new text features based on these markers and new polarity recalculation mechanisms based on their polarity and their distribution over text.

Notes

1. URL: http://extensions.openoffice.org/en/project/spanish-espanol.

2. URL: http://hdl.handle.net/10045/45506.

3. URL: http://lucene.apache.org/solr/.

4. URL: http://www.cs.waikato.ac.nz/ml/weka/.

5. URL: http://gplsi.dlsi.ua.es/attos/?q=node/110.

References

Arce, J. L. 2012. Medios de Comunicación de Masas en Costa Rica: Entre la digitalización, la convergencia y el auge de los "New Media". *Hacia la Sociedad de la Información y el Conocimiento, Programa Sociedad de la Información y el Conocimiento, Universidad de Costa Rica,* 283-308.

Cabanlit, Mark Anthony and Kurt Junshean Espinosa. 2014, July. Optimizing N-gram based text feature selection in sentiment analysis for commercial products in Twitter through polarity lexicons. *IISA 2014, The 5th International Conference on Information, Intelligence, Systems and Applications*, 94-97. IEEE.

Cambria, Erick *et al*. 2013. New avenues in opinion mining and sentiment analysis. *IEEE Intelligent Systems*, *28*(2), 15-21. Obtained from http://sentic.net/new-avenues-in-opinion-mining-and-sentiment-analysis.pdf

Chenlo, J. M. & Losada, D. E. 2014**.** An empirical study of sentence features for subjectivity and polarity classification. *Information Sciences, Elsevier,, 280*, 275-288.

Feldman, Ronen. 2013. Techniques and applications for sentiment analysis. *Communications of the ACM, 56*(4), 82-89. Obtained from http://dl.acm.org/citation.cfm?doid=2436256.2436274

Forman, G. 2003. An extensive empirical study of feature selection metrics for text classification *Journal of machine learning research, 3*, 1289-1305.

Guo, Liqiang and Wan, Xiaojun. 2012. Exploiting syntactic and semantic relationships between terms for opinion retrieval. *Journal of the American Society for Information Science and Technology, 63*(11), 2269-2282. Obtained from http://onlinelibrary.wiley.com/doi/10.1002/asi.22724/full

Indurkhya, N. & Damerau, F. J. 2010. Handbook of natural language processing *CRC Press, 2*.

Kouloumpis, E.; Wilson, T. & Moore, J. D. 2011. Twitter sentiment analysis: The good the bad and the omg. *Icwsm,11*, 538-541.

Martín-Valdivia, María Teresa *et al*. 2013. Sentiment polarity detection in Spanish reviews combining supervised and unsupervised approaches. *Expert Systems with Applications, 40*(10), 3934-3942. Obtained from http://www.sciencedirect.com/science/article/pii/S0957417412013267

Melero, M.; Cardús, A.-B.; Moreno, A.; Rehm, G.; de Smedt, K. & Uszkoreit, H. (2012). The Spanish language in the digital age. *Springer.*

Pang, Bo and Lee, Lillian. 2008. Opinion mining and sentiment analysis. *Foundations and trends in information retrieval, 2*(1-2), 1-135. Obtained from http://dx.doi.org/10.1561/1500000011

Perez-Rosas, Verónica *et al*. 2012, May. Learning Sentiment Lexicons in Spanish. *In LREC, 12,* 3077-3081.

Sharma, Anuj and Dey, Shubhamoy. 2012. Performance investigation of feature selection methods and sentiment lexicons for sentiment analysis. *IJCA Special Issue on Advanced Computing and Communication Technologies for HPC Applications, 3*, 15-20.

Stats. 2013. Internet World Users By Language: Top 10 Languages. Electronic site. Obtained from http://www.internetworldstats.com/stats7.htm

Turney, Peter D. 2002, July. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. *Proceedings of the 40th annual meeting on association for computational linguistics*, 417-424. Association for Computational Linguistics. Obtained from http://dl.acm.org/citation.cfm?doid=1073083.1073153