Article

A study on LIWC categories for opinion mining in Spanish reviews

Journal of Information Science
1–13
© The Author(s) 2014
Reprints and permissions:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/0165551510000000
jis.sagepub.com
SAGE

María del Pilar Salas-Zárate
Departamento de Informática y Sistemas,
Universidad de Murcia. Campus de Espinardo 30100 Murcia, Spain

Estanislao López-López
Departamento de Informática y Sistemas,
Universidad de Murcia. Campus de Espinardo 30100 Murcia, Spain

Rafael Valencia-García
Departamento de Informática y Sistemas,
Universidad de Murcia. Campus de Espinardo 30100 Murcia, Spain

Nathalie Aussenac-Gilles
CNRS ; IRIT ; Université Paul Sabatier (UPS) - 118 Route de Narbonne, F-31062 Toulouse Cedex 9, France

Ángela Almela
Department of Modern Languages, Universidad Católica San Antonio de Murcia, Spain

Giner Alor-Hernández
Division of Research and Postgraduate Studies Instituto Tecnológico de Orizaba, Av. Oriente 9, No. 852, Col. E. Zapata, CP 94320 Orizaba, Veracruz, Mexico

Abstract

With the exponential growth of social media i.e. blogs and social networks, organizations and individual persons are increasingly using the number of reviews of these media for decision making about a product or service. Opinion mining detects whether the emotions of an opinion expressed by a user on Web platforms in natural language, is positive or negative. This paper presents extensive experiments to study the effectiveness of the classification of Spanish opinions in five categories: highly positive, highly negative, positive, negative and neutral, using the combination of the psychological and linguistic features of LIWC. LIWC is a text analysis software that enables the extraction of different psychological and linguistic features from natural language text. For this study, two corpora have been used, one about movies and one about technological products. Furthermore, we have conducted a comparative assessment of the performance of various classification techniques: J48, SMO and BayesNet, using precision, recall and F-measure metrics. All in all, findings have revealed that the positive and negative categories provide better results than the other categories. Finally, experiments on both corpora indicated that SMO produces better results than BayesNet and J48 algorithms, obtaining an F-measure of 90.4% and 87.2% in each domain.

Keywords

Sentiment analysis; opinion mining; natural language processing with LIWC; machine learning

Corresponding author:
María del Pilar Salas-Zárate, Rafael Valencia-García, Departamento de Informática y Sistemas, Universidad de Murcia. Campus de Espinardo 30100 Murcia, Spain,
Email: mariapilar.salas@um.es, valencia@um.es.

1. Introduction

The dramatic spread of the Internet in society has substantially changed the forms of communication, entertainment, knowledge acquisition and consumption. There is a constant increase in the number of people who consider the Internet as a medium for answering their queries [1], in addition to using it as a powerful means of communication. Indeed, on the one hand, the reviews expressed in forums, blogs and social networks are having greater importance to make a decision to buy a product, hire a service, and vote for a political party, among others. On the other hand, for providers, this information is also important to get some feedback about their clients' expectations and needs, clients' feelings about their products or services and then to improve them. However, the number of reviews has increased exponentially on the Web, therefore reading all the opinions is impossible for the users. On these grounds, different technologies to automatically process these reviews have lately arisen. These technologies are usually known as opinion mining.

Sentiment analysis or opinion mining is a type of subjectivity analysis, which aims at identifying opinions, emotions and evaluations expressed in natural language. The main goal is to predict the sentiment orientation (i.e. positive, negative or neutral) of an evaluation by analysing sentiment or opinion words and expressions in sentences and documents. Three fundamental problems have to be solved which require at least linguistic (lexical and syntactical) language analysis, or a richer and formal text characterisation: aspect detection, opinion word detection and sentiment orientation identification [2]. The opinion mining task can be transformed into a classification task, so different supervised classification algorithms such as Support Vector Machines (SVM), Bayes Networks and Decision Trees can be used to solve this task.

Thanks to these techniques, several attempts at sentiment classification are being made. However, one of the main issues is that there are many conceptual rules that govern the linguistic expression of sentiments. Human psychology, which relates to social, cultural and other aspects, can be an important feature in sentiment analysis. For this reason, the sentiment mining process requires a rich and diverse text analysis as input. The LIWC text analysis software is a good candidate that enables the extraction of psychological and linguistic features from natural language text. We propose to evaluate how LIWC features can be used to classify reviews. It is worth noting that most of the studies on opinion mining deal exclusively with English and Chinese documents, perhaps owing to the lack of resources in other languages. Since the Spanish language has a much more complex syntax than many other languages, and is currently the third most spoken language in the world, we firmly believe that the computerization of Internet domains in this language is of utmost importance.

The aim of our work is to evaluate how the LIWC features can be used to classify Spanish reviews into five categories: positive, negative, neutral, highly positive or highly negative using different classifiers. For this purpose, two corpora of Spanish product reviews were first compiled. The first one is a corpus of movies, which has already been used in other studies. The second one is a corpus of technological products, which has been built from online selling websites. Secondly, the corpora were processed by LIWC to extract linguistic features. Then, three different classifying algorithms were evaluated on the processed corpora with the WEKA tool [3].

This paper is structured as follows: Section 2 presents the state of the art on opinion mining and sentiment analysis. Section 3 describes and discusses text analysis dimensions using LIWC. Section 4 presents the three classifiers used in WEKA for the experiment. Section 5 presents the evaluation of the classifiers based on LIWC text features and the classification of reviews into positive, negative, neutral, highly positive and highly negative. Also, a comparison of the results with related work is presented. Finally, Section 6 describes conclusions and future work.

2. Related Work

In recent years, several pieces of research have been conducted in order to improve sentiment classification. Many approaches [4, 5, 6, 7, 8, 9, 10, 11] proposed methods for the sentiment classification of English reviews.

For example, in [4] three corpora available for scientific research into opinion mining are analysed. Two of them are used in several studies, and the last one has been built ad-hoc from Amazon reviews on digital cameras. Finally, an SVM algorithm with different features is applied, in order to test how the sentiment classification is affected. The study presented in [5] proposes an empirical comparison between a neural network approach and an SVM-based method for classifying positive versus negative reviews. The experiments evaluate both methods as regards the function of selected terms in a bag-of-words (unigrams) approach. In [6] a comparative study of the effectiveness of ensemble methods for sentiment classification is presented. The authors consider two schemes of feature sets, three types of ensemble methods, and three ensemble strategies to conduct a range of comparative experiments on five widely-used datasets, with an emphasis on the evaluation of the effects of three ensemble strategies and the comparison of different ensemble methods. The results demonstrate that using an ensemble method is an effective way to combine different feature sets and classification algorithms for better classification performance. In this line of research, He, & Zhou [7] propose a novel framework where prior knowledge from a generic sentiment lexicon is used to build a classifier. The documents tagged by this classifier are used to automatically acquire domain-specific feature words, the word-class distributions of which are estimated and are subsequently used to train another classifier by constraining the model's predictions on unlabelled instances. The experiments, the movie-review data and the multi-domain sentiment dataset show that the approach attains comparable or better performance rates than existing hardly supervised sentiment classification methods despite using no labelled documents. In [8] the authors propose an innovative methodology for opinion mining that brings together traditional natural language processing techniques with sentiment analysis processes and Semantic Web technologies. The aim of this work is to improve feature-based opinion mining by employing ontologies in the selection of features and to provide a new method for sentiment analysis based on vector analysis. In [9] a comparative study among n-grams (unigram, bigram and trigram) method and feature weighting (TF and TF-IDF) is presented. In this piece of research, messages of Twitter to review a movie are used for opinion mining. Also, this work is only related to sentiment classification into two classes (binary classification), that is, a positive class and negative class. The positive class shows good message opinion; otherwise, the negative class shows the bad message opinion of certain movies. The study presented in [10] proposes a new unsupervised approach to the problem of polarity classification in Twitter posts. The polarity classification problem is resolved by combining SentiWordNet scores with a random walk analysis of the concepts found in the text over the WordNet graph. In order to validate their unsupervised approach, several experiments were performed in order to analyse major issues in the method and to compare it with other approaches like plain SentiWordNet scoring or machine learning solutions such as Support Vector Machines in a supervised approach. Chen, Liu, & Chiu [11] propose a neural-network based approach. It uses semantic orientation indexes as input for the neural networks to determine the sentiments of the bloggers quickly and effectively. Several blogs are used to evaluate the effectiveness of the approach. The results indicate that the proposed approach outperforms traditional ones including other neural networks and several semantic orientation indexes.

Furthermore, other proposals [12, 13, 14, 15] introduce methods for sentiment classification of Chinese reviews. Zhai, Xu, & Jia [12] analyze sentiment-word, substring, substring-group, and key-substring-group features, and the commonly used Ngram features. To explore general language, two authoritative Chinese datasets in different domains were used. The statistical analysis of the results indicates that different types of features possess different discriminative capabilities in Chinese sentiment classification. Xu, Peng, & Cheng [13] propose a new method for identifying the semantic orientation of subjective terms to perform sentiment analysis. The method takes a classification approach that is based on a novel semantic orientation representation model called S-HAL (Sentiment Hyperspace Analogue to Language). The results indicate that this method has outperformed the SO-PMI method and several other published methods. In [14] a two-stage framework for cross-domain sentiment classification is proposed. A bridge between the source domain and the target domain is built with the aim of getting some of the most reliably labelled documents in the target domain. The results indicate that the proposed approach could improve the performance of cross-domain sentiment classification dramatically. In [15] a study presents the standpoint that uses individual model (i-model) based on artificial neural networks (ANNs) to determine text sentiment classification. The individual model comprises sentimental features, feature weight and prior knowledge base. The results of the experiment show that the accuracy of the individual model is higher than that of support vector machines (SVMs) and hidden Markov model (HMM) classifiers on the movie review corpus.

Finally, it is worth noting that not many proposals such as the one presented here [16] are focused on sentiment classification of Spanish reviews. In this work, two lexicons are used to classify the opinions using a simple approach based on counting the number of words included in the lexicons that occur in each evaluation. Specifically, an opinion is positive if the number of positive words is greater than or equal to the number of negative ones, and is negative in the opposite case.

In order to fully analyse the studies described above and compare them with our proposal, a comparative table is provided below (see Table 1) which summarizes relevant properties of these pieces of research. For this comparison, four features have been used: 1) computational learning, 2) linguistic resources, 3) domain, and 4) language.

Several machine learning techniques are used, i.e. SVM, Naïve Bayes, among others. Almost all the proposals use computational learning. Specifically, the SVM technique is the most frequently used [4, 5, 6, 7, 9, 10, 12, 13, 15]. Besides, the techniques of Naïve Bayes [6, 7, 10] and neural networks [11] are also used. On the other hand, other pieces of research do not use any machine learning technique [8, 14, 16].

The techniques used for polarity detection in these approaches are n-grams [4, 6, 9, 12, 15], term frequency [4, 6, 9, 12], and semantic orientation indexes [11]. Alternative approaches only use lexical resources [16].

Almost all the corpora used in the proposals mentioned above include reviews on movies [4, 5, 6, 8, 11, 15, 16]. Other proposals use corpora that include reviews on topics such as: music [11], hotels [4, 12], products [12, 14], news [13], DVDs [6, 7] and electronics [7].

The English language is the most used in these studies [4, 5, 6, 7, 8, 9, 10, 11]. However, other languages are used in some proposals, such as Chinese [12, 13, 14, 15] and Spanish [16].

On the basis of the results obtained from the comparative analysis summarized in Table 1, the present study seeks to evaluate the performance of three different classifying algorithms in the classification of Spanish opinions through the combination of psychological and linguistic features extracted using the LIWC text analyser.

Table 1. Comparison of proposals for sentiment classification.

Proposal	Computational learning	Linguistic resources	Domain	Language
[4]	Yes (SVM(Support Vector Machines))	Ngrams, TF-IDF (Term frequency – Inverse document frequency), BO (Binary Occurrence) and TO (Term Occurrence)	Movies, books, cars, cookware, hotels, music, cameras, phones and computers	English
[5]	Yes (SVM and ANN(Artificial Neural Network))	Bag-of-words model	Movies, GPS, books and cameras	English
[6]	Yes (NB (Naïve Bayes), ME (maximum entropy) and SVM)	Ngrams and TF-IDF	Movie, books, DVDs, Electronics and Kitchen.	English
[7]	Yes (NB, SVM and ME)	Sentiment dictionary	Books, DVDs, electronics and Kitchen.	English
[8]	No	Sentiment dictionary(SentiwordNet) and ontologies	Movies	English
[9]	Yes (SVM)	Ngrams, TF and TF-IDF	Twitter (movies)	English
[10]	Yes (SVM, NB and ME)	SentiwordNet	Twitter (politics, business, economics)	English
[11]	Yes (NN(Neural Network))	BPN (back-propagation neural network) and SO indexes.	Movie, MP3 and Blog	English and Chinese
[12]	Yes (SVM)	Ngrams and TFIDF—C	Hotels and products	Chinese
[13]	Yes (SVM)	Sentiment dictionary(S-HAL)	News	Chinese
[14]	No	Sentiment dictionary (SentiRank)	Products	Chinese
[15]	Yes (ANN and SVM)	Ngrams	Movie	Chinese
[16]	No	BLEL: the Bing Liu English Lexicon	Movie	Spanish and English

3. LIWC

LIWC (Linguistic Inquiry and Word Count) is a software application that provides an effective tool for studying the emotional, cognitive, and structural components contained in language on a word-by-word basis. Early approaches to psycholinguistic concerns involved almost exclusively qualitative philosophical analyses. More modern research in this field provides empirical evidence on the relation between language and the state of mind of subjects, or even their mental health [17]. In this regard, further studies such as [18] have dealt with the therapeutic effect of verbally expressing emotional experiences and memories. LIWC was developed precisely for providing an efficient method for studying these psycholinguistic concerns thanks to corpus analysis, and has been considerably improved since its first version [19]. An updated revision of the original application was presented in [20], namely LIWC2001.

LIWC provides a Spanish dictionary composed by 7,515 words and word stems. Each word can be classified into one or more of the 72 categories included by default in LIWC. Also, the categories are classified into four dimensions: (1) standard linguistic processes, (2) psychological processes, (3) relativity, and (4) personal concerns.

Next, Table 2 shows some examples of the LIWC categories. The full list of categories is presented in [21].

Table 2. LIWC categories

1. Linguistic processes	Word count, total pronouns, articles, prepositions, numbers, negations
2. Psychological Processes	Affective process, positive emotions, negative emotion, cognitive process, perceptual process
3. Relativity	Time, space, motion
4. Personal concerns	Occupation, leisure activity, money/financial issues, religion, death and dying

As can be seen in Table 2, the first dimension, standard linguistic processes, involves function words and grammatical information, whereas the second and fourth dimensions are more subjective, especially those denoting emotional processes within the second dimension. Within this dimension, the emotion or affective processes are using sub-dictionaries which gather words selected from several sources such as the PANAS [22] and Roget's Thesaurus, being subsequently rated by groups of three judges working independently. Similar to the first dimension, the third dimension, "relativity", is composed of a category concerning time, which is quite clear: past, present, and future tense verbs. Within the same dimension, the space category includes spatial prepositions and adverbs . Finally, the fourth dimension involves word categories related to personal concerns intrinsic to the human condition. This is important because it can affect the voicing of a feeling in an opinion.

4. Data Sets

For the present study, a set of reviews in Spanish that include positive, negative, neutral, highly positive and highly negative reviews was necessary. Each review text is assigned to a single category, meaning that the review as a whole is either positive, negative, etc. Therefore, two corpora were collected, one within the domain of product reviews and the other one within the domain of movie reviews. The first one contains 600 reviews of technological products such as mobile devices, specifically 100 highly negative reviews, 150 negative reviews, 100 neutral reviews, 150 positive reviews and 100 highly positive reviews, obtained from online selling websites e.g. moviles.com [23]. Also, each review was examined and classified manually to ensure its quality. The second corpus was obtained from the corpus presented in [24] related to movie reviews. The original corpus contains 3,878 opinions, which are already classified into five categories (351 highly negative reviews, 923 negative reviews, 1,253 neutral reviews, 890 positive reviews and 461 highly positive reviews). For this experiment, a corpus of 1,000 opinions was compiled by selecting 200 random opinions for each category.

Once the corpora have been built, they are analysed through all the possible combinations of LIWC dimensions and taking into account three possible sets of opinion classes (positive-negative, positive-neutral-negative and highly positive-positive-neutral-negative-highly negative). LIWC searches for target words or word stems from the dictionary, categorizes them into one of its linguistic dimensions, and then converts the raw counts to percentages of total words. The values obtained for the categories were used for the subsequent training of the machine learning classifier.

This analysis aims to evaluate the classifying potential of these dimensions, both individually and collectively.

It is worth noting that the results obtained by LIWC were manually evaluated by experts to confirm that LIWC produces correct results when analysing a set of reviews.

5. Machine Learning and classification

In the present work, WEKA [3] has been used to evaluate the classification success of reviews (positive, negative, neutral, highly positive or highly negative) based on LIWC categories.

Table 3. Machine learning methods

Classifier	Description
J48	J48 was developed by Ross Quinlan. It is an open source Java implementation of the C4.5 algorithm in the Weka data mining tool. The algorithm uses an advanced technique to induce decision trees for classification and uses reduced-error pruning. The purpose of classifying data with decision tree is to discover if it contains well-separated classes of items that can be interpreted meaningfully [25]
BayesNet	Bayesian networks are directed acyclic graphs in which the nodes represent propositions (or variables), the arcs imply the existence of direct causal dependencies between the linked propositions, and the strengths of these dependencies are quantified by conditional probabilities. Such a network can be used to represent the deep causal knowledge of an agent or a domain expert and turns into a computational architecture if the links are used not merely for storing factual knowledge, but also for directing and activating the data flow in the computations which manipulate this knowledge [26].
SMO	Sequential minimal optimization (SMO) was developed by John Platt in 1998. SMO is an improved training algorithm for SVMs. Like other SVM training algorithms, SMO breaks down a large QP (quadratic programming) problem into a series of smaller QP problems. Unlike other algorithms, SMO utilizes the smallest possible QP problems, which are solved quickly and analytically, generally improving its scaling and computation time significantly [27].

WEKA provides several classifiers, which allows the creation of models according to the data and purpose of analysis. Classifiers are categorized into seven groups: Bayesian (Naïve Bayes, Bayesian nets, etc.), functions (linear regression, SMO, logistic, etc.), lazy (IBk, LWL, etc.), meta-classifiers (Bagging, Vote, etc.), miscellaneous (SerializedClassifier and InputMappedClassifier), rules (DecisionTable, OneR, etc.) and trees (J48, RandomTree, etc.). The classification process involves the building of a model based on the analysis of the instances. This model is represented through classification rules, decision trees, or mathematical formulae. The model is used to generate the classification of unknown data, calculating the percentage of instances which were correctly classified.

The experiment has been performed by using three different algorithms: the C4.5 decision tree (J48), the Bayes Network learning algorithm (BayesNet) and the SMO algorithm for SVM classifiers [28]. These algorithms were selected because they have been used in several experiments obtaining good results in data classification [29], [30].

Next, a brief description of the machine learning methods chosen for evaluation is presented in Table 3.

6. Evaluation and Results

6.1. Results and figures for the technological corpus

In order to evaluate the results of the classifiers, we have used three metrics: precision, recall and F-measure. Recall is the proportion of actual positive cases that were correctly predicted as such. On the other hand, precision represents the proportion of predicted positive cases that are real positives. Finally, F-measure is the harmonic mean of precision and recall.

For each classifier, a ten-fold cross-validation has been done. This technique is used to evaluate how the results obtained would generalise to an independent data set. Since the aim of this experiment is the prediction of the positive, negative, neutral, highly positive and highly negative condition of the texts, a cross-validation is applied in order to estimate the accuracy of the predictive models. It involves partitioning a sample of data into complementary subsets, performing an analysis on the training set and validating the analysis on the testing or validation set.

Next, the results of precision (P), recall (R), and the F-measure for each algorithm are reported (table 4-9). The first column indicates which LIWC dimensions are used, i.e. 1) standard linguistic processes, 2) psychological processes, 3) relativity, and 4) personal concerns.

The tables below show the results obtained for the classification of technological product reviews by using two, three and five categories: positive-negative (see Table 4), positive-neutral-negative (see Table 5) and highly positive-positive-neutral-negative-highly negative (see Table 6). In the first column, the number of LIWC dimensions used for each classifier is indicated. For example, 1_2_3_4 indicates that all the dimensions have been used in the experiment, and 1_2 indicates that only the categories of dimensions 1 and 2 have been used to train the classifier.

Table 4. Products with two categories (positive and negative).

	J48			BayesNet			SMO		
	P	R	F1	P	R	F1	P	R	F1
1	0.739	0.741	0.74	0.799	0.797	0.797	0.843	0.843	0.843
2	0.799	0.8	0.799	0.833	0.833	0.833	0.823	0.822	0.822
3	0.733	0.735	0.73	0.781	0.782	0.781	0.796	0.795	0.795
4	0.742	0.741	0.741	0.76	0.761	0.761	0.755	0.755	0.755
1_2	0.803	0.804	0.803	0.885	0.881	0.882	0.887	0.887	0.886
1_3	0.75	0.752	0.751	0.819	0.818	0.819	0.832	0.833	0.832
1_4	0.771	0.771	0.771	0.813	0.811	0.812	0.832	0.833	0.832
2_3	0.819	0.82	0.819	0.879	0.878	0.878	0.863	0.863	0.863
2_4	0.808	0.809	0.809	0.854	0.852	0.853	0.845	0.845	0.844
3_4	0.737	0.737	0.737	0.811	0.811	0.811	0.817	0.818	0.817
1_2_3	0.816	0.816	0.816	0.889	0.885	0.886	0.881	0.881	0.881
1_2_4	0.821	0.822	0.82	0.87	0.865	0.867	0.879	0.879	0.879
1_3_4	0.803	0.804	0.802	0.828	0.827	0.828	0.837	0.838	0.837
2_3_4	0.805	0.806	0.804	0.878	0.874	0.875	0.866	0.867	0.867
1_2_3_4	0.83	0.831	0.83	0.878	0.874	0.875	0.904	0.905	**0.904**

Table 5. Products with three categories (positive-neutral-negative).

	J48			BayesNet			SMO		
	P	R	F1	P	R	F1	P	R	F1
1	0.677	0.687	0.682	0.697	0.687	0.692	0.743	0.745	0.744
2	0.671	0.669	0.670	0.703	0.709	0.706	0.713	0.732	0.722
3	0.605	0.634	0.619	0.597	0.645	0.620	0.585	0.678	0.628
4	0.614	0.622	0.618	0.614	0.66	0.636	0.561	0.649	0.602
1_2	0.705	0.704	0.704	0.769	0.753	0.761	0.772	0.782	0.777
1_3	0.736	0.747	0.741	0.775	0.777	0.776	0.714	0.707	0.710
1_4	0.672	0.68	0.676	0.717	0.71	0.713	0.717	0.727	0.722
2_3	0.693	0.706	0.699	0.745	0.75	0.747	0.732	0.75	0.741
2_4	0.667	0.675	0.671	0.739	0.742	0.740	0.73	0.744	0.737
3_4	0.652	0.658	0.655	0.661	0.695	0.678	0.716	0.713	0.714
1_2_3	0.677	0.678	0.677	0.763	0.748	0.755	0.776	0.785	0.780
1_2_4	0.698	0.704	0.701	0.774	0.759	0.766	0.769	0.78	0.774
1_3_4	0.665	0.672	0.668	0.727	0.719	0.723	0.738	0.748	0.743
2_3_4	0.688	0.692	0.690	0.759	0.759	0.759	0.753	0.771	0.762
1_2_3_4	0.678	0.686	0.682	0.766	0.753	0.759	0.774	0.786	0.780

Table 6. Products with five categories (highly positive-positive-neutral-negative-highly negative).

	J48			BayesNet			SMO		
	P	R	F1	P	R	F1	P	R	F1
1	0.407	0.413	0.41	0.433	0.463	0.447	0.473	0.506	0.489
2	0.458	0.468	0.462	0.479	0.501	0.49	0.447	0.494	0.469
3	0.394	0.398	0.395	0.328	0.445	0.376	0.364	0.471	0.410
4	0.375	0.381	0.377	0.357	0.45	0.397	0.467	0.456	0.461
1_2	0.495	0.497	0.496	0.517	0.526	0.521	0.536	0.543	0.539
1_3	0.417	0.422	0.419	0.445	0.469	0.457	0.455	0.54	0.493
1_4	0.426	0.422	0.424	0.457	0.475	0.466	0.48	0.514	0.496
2_3	0.496	0.5	0.498	0.514	0.532	0.523	0.475	0.518	0.495
2_4	0.478	0.479	0.478	0.504	0.526	0.515	0.473	0.509	0.49
3_4	0.425	0.421	0.422	0.393	0.443	0.416	0.466	0.495	0.48
1_2_3	0.464	0.469	0.466	0.515	0.523	0.519	0.528	0.544	0.536
1_2_4	0.498	0.498	0.498	0.517	0.529	0.523	0.525	0.535	0.53
1_3_4	0.426	0.421	0.423	0.449	0.477	0.463	0.486	0.526	0.505
2_3_4	0.452	0.454	0.452	0.518	0.538	0.528	0.487	0.518	0.502
1_2_3_4	0.514	0.512	0.513	0.526	0.538	0.532	0.567	0.575	0.571

Considering the tables above, the different classification algorithms generally show similar results, although SVM obtains better results. The best classification results were obtained using two categories, positive-negative (see Table 4). Also, the results from the J48 algorithm show that individually, the second dimension, "psychological processes", provides the best results, with an F-measure of 79.9%. Conversely, the third dimension, "relativity", provides the worst results, with an F-measure of 73.0%. On the other hand, the combination of all LIWC dimensions provides the best classification result with an F-measure of 83.0%.

The results from the BayesNet algorithm are similar to the ones obtained by the J48 algorithm, although this experiment provides better classification results. The results show that the second dimension, "psychological processes", provides the best results on its own as well, with an F-measure of 83.3%. Quite the reverse, the fourth dimension, "personal concerns", provides the worst results with an F-measure of 76.1%. Furthermore, the combination of 1_2_3 LIWC dimensions provides the best classification result, with an F-measure of 88.6%. The results obtained by means of the use of the four dimensions are also good, with an overall F-measure of 87.5%

The results from the experiment with SMO are better than the ones obtained with the previous algorithms. The results show that, once again, the first dimension provides the best results by itself, with an F-measure of 84.3%. On the contrary, the fourth dimension, "personal concerns", provides the worst results with a score of 75.5%. Moreover, the combination of all LIWC dimensions provides the best classification result, with an F-measure of 90.4%.

6.2. Results and figures for the movies corpus

The tables below show the results obtained for the classification of movie reviews by using two, three and five categories: positive-negative (see Table 7), positive-neutral-negative (see Table 8) and highly positive-positive-neutral-negative-highly negative(see Table 9).

In the classification results for the corpus of movies (Table 7, Table 8 and Table 9), we found that BayesNet algorithm (Table 7) gets the best results using two categories (positive-negative). When considering the results from the J48 algorithm, they show that individually, the first dimension, "standard linguistic processes", provides the best results, with an F-measure of 77.3%. Quite the reverse, the fourth dimension, "personal concern", provides the worst results, with an F-measure of 68.2%. In addition, the combination of 1_2_3 LIWC dimensions provides the best classification result with an F-measure of 79.6%.

The results from the experiment with BayesNet algorithm provides better classification results than J48 algorithm. The results show that the first dimension, "standard linguistic processes", provides the best results on its own as well, with an F-measure of 81.3%. Conversely, the fourth dimension, "personal concerns", provides the worst results with an F-measure of 68.2%. Besides, the combination of 1_2 LIWC dimensions provides the best classification result, with an F-measure of 82.8%.

The results of the SMO algorithm are even better than the ones obtained with the two previous algorithms. The results show that, once again, the first dimension, "standard linguistic processes", provides the best results by itself, with an F-measure of 81.1%. On the contrary, the fourth dimension, "personal concerns", provides the worst results with a score of 73.8%. Furthermore, the combination of 1_2_4 LIWC dimensions provides the best classification result with an F-measure of 87.2%.

Table 7. Movies with opinion classification (positive and negative).

	J48			BayesNet			SMO		
	P	R	F1	P	R	F1	P	R	F1
1	0.774	0.773	0.773	0.814	0.813	0.813	0.812	0.811	0.811
2	0.711	0.711	0.711	0.721	0.721	0.721	0.796	0.796	0.796
3	0.755	0.753	0.753	0.705	0.703	0.703	0.745	0.744	0.744
4	0.682	0.682	0.682	0.682	0.682	0.682	0.738	0.738	0.738
1_2	0.78	0.78	0.78	0.828	0.828	0.828	0.869	0.868	0.868
1_3	0.769	0.769	0.769	0.811	0.81	0.81	0.827	0.827	0.827
1_4	0.764	0.763	0.763	0.808	0.806	0.806	0.834	0.834	0.834
2_3	0.684	0.684	0.684	0.745	0.745	0.745	0.808	0.808	0.808
2_4	0.721	0.721	0.721	0.722	0.722	0.722	0.805	0.805	0.805
3_4	0.728	0.728	0.728	0.719	0.719	0.719	0.754	0.754	0.754
1_2_3	0.796	0.796	0.796	0.822	0.82	0.82	0.872	0.871	0.871
1_2_4	0.781	0.781	0.781	0.819	0.819	0.819	0.872	0.872	**0.872**
1_3_4	0.754	0.754	0.754	0.817	0.816	0.816	0.844	0.844	0.844
2_3_4	0.711	0.711	0.711	0.745	0.745	0.745	0.819	0.819	0.819
1_2_3_4	0.778	0.778	0.778	0.82	0.819	0.819	0.855	0.854	0.854

Table 8. Movies with three categories

	J48			BayesNet			SMO		
	P	R	F1	P	R	F1	P	R	F1
1	0.558	0.565	0.561	0.648	0.662	0.654	0.569	0.678	0.618
2	0.56	0.564	0.562	0.574	0.599	0.586	0.565	0.672	0.613
3	0.544	0.565	0.554	0.494	0.583	0.534	0.516	0.613	0.560
4	0.54	0.544	0.542	0.44	0.537	0.483	0.513	0.609	0.556
1_2	0.581	0.581	0.581	0.682	0.69	0.685	0.687	0.714	0.7
1_3	0.565	0.572	0.568	0.661	0.65	0.655	0.567	0.675	0.616
1_4	0.582	0.574	0.577	0.644	0.659	0.651	0.584	0.697	0.635
2_3	0.571	0.571	0.571	0.614	0.619	0.616	0.641	0.687	0.663
2_4	0.567	0.57	0.568	0.566	0.589	0.577	0.605	0.665	0.633
3_4	0.517	0.518	0.517	0.491	0.58	0.531	0.536	0.636	0.581
1_2_3	0.584	0.582	0.583	0.669	0.675	0.672	0.696	0.721	0.708
1_2_4	0.594	0.588	0.591	0.68	0.688	0.683	0.677	0.712	0.694
1_3_4	0.584	0.585	0.584	0.641	0.655	0.648	0.621	0.693	0.655
2_3_4	0.557	0.562	0.559	0.613	0.617	0.615	0.637	0.672	0.654
1_2_3_4	0.582	0.579	0.581	0.67	0.676	0.672	0.697	0.724	0.71

Table 9. Movies with five categories.

	J48			BayesNet			SMO		
	P	R	F1	P	R	F1	P	R	F1
1	0.414	0.415	0.414	0.465	0.482	0.473	0.498	0.508	0.503
2	0.522	0.527	0.524	0.447	0.445	0.446	0.439	0.434	0.436
3	0.397	0.397	0.397	0.37	0.368	0.369	0.421	0.43	0.425
4	0.339	0.34	0.339	0.264	0.343	0.298	0.366	0.385	0.375
1_2	0.403	0.403	0.403	0.462	0.482	0.472	0.538	0.547	0.542
1_3	0.406	0.408	0.407	0.454	0.47	0.462	0.497	0.503	0.5
1_4	0.418	0.421	0.419	0.465	0.484	0.474	0.501	0.512	0.506
2_3	0.375	0.375	0.375	0.402	0.43	0.415	0.482	0.485	0.483
2_4	0.391	0.391	0.391	0.403	0.443	0.422	0.462	0.472	0.467
3_4	0.377	0.377	0.377	0.325	0.368	0.345	0.418	0.43	0.424
1_2_3	0.424	0.422	0.423	0.475	0.485	0.48	0.533	0.54	0.536
1_2_4	0.419	0.419	0.419	0.465	0.486	0.475	0.523	0.531	0.526
1_3_4	0.394	0.394	0.394	0.451	0.472	0.461	0.5	0.507	0.503
2_3_4	0.373	0.372	0.372	0.401	0.427	0.413	0.474	0.483	0.478
1_2_3_4	0.428	0.425	0.426	0.477	0.487	0.482	0.522	0.525	0.523

6.3. Discussion of the results

General results show that the combination of different LIWC dimensions provides better results than individual dimensions. Individually, the first one and the second one provides the best results, probably due to the great amount of grammatical words that are part of the standard linguistic dimension and the fact that written opinions frequently contain words related to the emotional state of the author containing word stems classified into categories such as anxiety, sadness, positive and negative emotions, optimism and energy, and discrepancies, among others. All these categories are included in the second dimension, confirming its discriminatory potential in classification experiments. Furthermore, the high performance of the first dimension is natural, bearing in mind the considerable potential of function words, which constitutes a substantial part of standard linguistic dimensions. The prime importance of these grammatical elements has been widely explored, not only in computational linguistics, but also in psychology. As Chung and Pennebaker (2007: 344) have it, these words "can provide powerful insight into the human psyche". Variations in their usage have been associated to sex, age, mental disorders such as depression, status, and deception [31]. On the other hand, the fourth dimension provides the worst results, owing to the fact that the topics selected for this study, "technological products" and "movies", bears little relation to the vocabulary corresponding to "personal concerns" categories. It can be stated that this dimension is the most content-dependent, and thus the least revealing.

As regards the classification with two categories (positive-negative), it provides better results than the classification with three (positive-neutral-negative) and five (highly positive-positive-neutral-negative-highly negative) categories. Thus, it is by virtue of the combination of fewer categories that the classification algorithm performs a better classification, probably due to the fact that in a bipolar system there is less space for the classification of slippery cases. It also means that additional criteria and features are required to get a fine-grained classification into 5 categories for instance.

The results obtained for different classifiers are similar. However, SMO provides better results than J48 and BayesNet. These results can be justified by the analysis of different algorithms present in [32], where it is clearly shown how SVM models are more robust and accurate compared to other classifiers, including the ones used in this piece of research. Furthermore, SVMs have been successfully applied to many text classification tasks due to their main advantages: first, they are robust in high dimensional spaces; second, any feature is relevant; third, they are robust when there is a sparse set of samples; and finally, most text categorization problems are linearly separable [4]. Unlike other classifiers such as decision trees or logistic regressions, SVM assumes no linearity, and it can be difficult to interpret its results outside its accuracy values [33].

Finally, with regard to the classification results for the corpus of movie reviews, they are worse than those for the corpus of technological products. From our point of view, the classification results through the LIWC dimensions are strongly dependent on the topic. Thus, for example, the combination of 1_2_3_4 (90.4%) LIWC dimensions achieves the best results for the corpus of "technological products". In contrast, 1_2_4 (87.2%) LIWC dimensions show the best result for the corpus of "movies". There is no doubt that the factor loadings of the four dimensions play a considerable part here.

6.4. Comparison with related work

As stated earlier, many different approaches exist for sentiment classification, and opinion analysis in English and Chinese. Moreover, the results from most of the approaches in these languages present better results than other proposals for Spanish language. We considered that the interest in the English language arises from the fact that it is an official language in a large number of countries, and most of the content on the Internet is written in this language. As regards Chinese, it is becoming one of the most important languages for international business. As commented on in Section 2, extensive research has been carried out for these languages, but not all of them have been evaluated using the standard measures. Thus, Table 10 shows the results from those studies that have been evaluated in terms of precision, recall and F1.

Table 10. Comparison of related work with our proposal.

Proposal	Language	Precision	Recall	F1
[4]	English	84.01	85.80	84.79
[5]	English	84.00	87.00	85.47
[7]	English	N/A	N/A	83.26
[9]	English	68.81	81.87	74.77
[10]	English	64.29	61.47	62.85
[11]	English and Chinese	75.00	61.90	67.80
[12]	Chinese	86.40	87.60	86.90
[13]	Chinese	N/A	N/A	92.21
[16]	Spanish and English	63.93	62.74	63.33
Our proposal	Spanish	90.4	90.5	90.4
		87.2	87.2	87.2

Table 10 shows that our proposal obtained similar results to other approaches, with a high F-measure of 90.4% and 87.2%. However, it is difficult to compare the different opinion mining approaches described in the literature, because none of the software applications is available. Indeed, the corpora used for each experiment differ significantly in content and size, topics and language. A fair comparison of two opinion mining methods would require the usage of the same testing corpus. In spite of this, Table 10 shows that in [16] the system obtains an F-measure of 63.33% for the Spanish language, which is considerably lower than the value obtained by our approach.

The studies for the English and Chinese language obtained similar F-measures to the ones obtained here. For example, proposals in English [4] and [5] obtained F-measures of 84.79% and 85.47%, and proposals in Chinese [12] and [13] obtained F-measures of 86.90% and 92.91%, respectively. However, it is important to mention the lower level of grammatical complexity of the English and Chinese languages as compared to Spanish, which seems to have a strong impact on the final results. For example, in Chinese, there are no tenses and conjugations for every verb.

7. Conclusions and Future Work

In this piece of research, we have presented an experiment based on sentiment classification with the aim of evaluating the classifying potential of LIWC dimensions. In order to conduct a comprehensive study, we have considered two, three and five categories: positive-negative, positive-neutral-negative and highly positive-positive-neutral-negative-highly negative for the classification of reviews in Spanish. Subsequently, in an attempt to evaluate the efficacy of LIWC features, J48, BayesNet and SMO Weka classifiers have been used. The results show that the classification of reviews with two categories "positive-negative" provides better results than with other categories. Also, SMO is a classifier that has obtained the best classification results. Finally, regarding the comparison with the related work, our proposal has obtained encouraging results with a high F-measure score of 90.4% for the corpus of technological product reviews and 87.2% for the corpus of movie reviews.

Despite all the advantages and possibilities of the proposed approach, it has several limitations that could be improved in future work. First, our approach lacks robustness due to the fact that all the input to LIWC must be grammatically correct. Furthermore, LIWC presents limitations of disambiguation and ignores context, irony, sarcasm, and idioms [34]. Second, our approach does not make use of other sentiment analysis techniques based on sentiment lexicons such as SentiWordNet [35]. Finally, our approach obtains the global polarity of a review. This is a drawback, because an entire document or a single sentence could contain different opinions about different features of the same product or service [36]. In fact, classifying opinions at the document or sentence level does not indicate what the user likes and dislikes. A positive report on an object does not mean that the user has positive opinions on all aspects or features of that object. Likewise, it would be inaccurate to state that a negative document entails that the user dislikes everything about the object. In a document (e.g., a product review), the user typically writes about both the positive and negative aspects of the object, although the general sentiment toward that object may be positive or negative [37]. To obtain such detailed aspects, it is necessary to perform feature-based opinion mining in an attempt to identify the features in the opinion and to classify the sentiments of the opinion for each of these features [38].

As regards further research, the authors are considering a new corpus where the vocabulary is better aligned with the "personal concerns" dimension, as well as other new corpora comprising different domains of the Spanish language, since research into sentiment classification in this language is needed. Furthermore, we will use LIWC features in English and French to verify whether this technique can be applied to different languages. On the other hand, we also attempt to apply the Probabilistic Latent Semantic Indexing to automated document indexing. Finally, it is also intended to adapt this approach to a feature-based opinion mining guided by ontologies, as in the study presented in [8].

Funding

This work has been partially supported by the Spanish Ministry of Economy and Competitiveness and the European Commission (FEDER / ERDF) through project SeCloud (TIN2010-18650). María del Pilar Salas-Zárate is supported by the National Council of Science and Technology (CONACYT), the Public Education Secretary (SEP) and the Mexican government. Additionally, this work has been supported by the University Paul Sabatier under its visiting professors programme.

References

[1] García-Crespo A, Colomo-Palacios R, Gómez-Berbís JM, and Ruiz-Mezcua B. SEMO: a framework for customer social networks analysis based on semantics. Journal of Information Technology, 2010; 25(2): 178-188.

[2] Thet TT, Cheon J, and Khoo C. Aspect-based sentiment analysis of movie reviews on discussion boards. Journal of Information Science, 2010; 36: 823-848.

[3] Bouckaert R, Frank E, Hall M, Holmes G, Pfahringer B, Reutemann P, and Witten I. WEKA—Experiences with a Java Open-Source Project. Journal of Machine Learning Research, 2010; 11: 2533-2541.

[4] Rushdi Saleh M, Martín Valdivia M, Montejo Ráez A, and Ureña López L. Experiments with SVM to classify opinions in different domains. Expert Systems with Applications, 2011; 38: 14799-14804.

[5] Moraes R, Valiati J, and Gavião Neto W. Document-level sentiment classification: An empirical comparison between SVM and ANN. Expert Systems with Applications, 2013; 40: 621-633.

[6] Xia R, Zong C, and Li S. Ensemble of feature sets and classification algorithms for sentiment classification. Information Sciences, 2011; 181: 1138-1152.

[7] He Y, and Zhou D. Self-training from labeled features for sentiment analysis. Information Processing and Management, 2011; 47: 606-616.

[8] Peñalver Martínez I, Valencia García R, and García Sánchez F. Ontology-guided approach for Feature-Based Opinion Mining. In: 16th International Conference on Applications of Natural Language to Information Systems, NLDB, 2011. Alicante, Spain.

[9] Basari SH, Hussin B, Ananta GP, and Zeniarja J. Opinion Mining of Movie Review using Hybrid Method of Support Vector Machine and Particle Swarm Optimization. Procedia Engineering, 2013; 53: 453-462.

[10] Montejo Ráez A, Martínez Cámara E, Martín Valdivia MT, and Ureña López LA. Ranked WordNet graph for Sentiment Polarity Classification in Twitter. Computer Speech and Language, 2014; 28: 93-107.

[11] Chen LS, Liu CH., and Chiu HJ. A neural network based approach for sentiment classification in the blogosphere. Journal of Informetrics, 2011, 5: 313-322.

[12] Zhai Z, Xu H, Kang B, and Jia P. Exploiting effective features for Chinese sentiment classification. Expert Systems with Applications, 2011; 38: 9139-9146.

[13] Xuo T, Peng Q, and Cheng Y. Identifying the semantic orientation of terms using S-HAL for sentiment analysis. Knowledge-Based Systems, 2012; 35: 279-289.

[14] Wu Q, and Tan S. A two-stage framework for cross-domain sentiment classification. Expert Systems with Applications, 2011; 38: 14269-14275.

[15] Jian Z, Chen X, and Han-Shi W. Sentiment classification using the theory of ANNs. The Journal of China Universities of Posts and Telecommunications, 2010; 17: 58-62.

[16] Molina González M, Martínez Cámara E, Martín Valdivia M, and Perea Ortega J. Semantic orientation for polarity classification in Spanish reviews. Expert Systems with Applications, 2013; 40: 7250-7257.

[17] Stiles WB. Describing Talk: A Taxonomy of Verbal Response Modes. Newbury Park, CA: Sage, 1992.

[18] Pennebaker JW, Francis ME, and Mayne TJ. Linguistic Predictors of Adaptive Bereavement. Journal of Personality and Social Psychology, 1997; 72(4): 863-871.

[19] Francis ME, and Pennebaker JW. LIWC: Linguistic Inquiry and Word Count. Dallas, TX: Southern Methodist University, 1993.

[20] Pennebaker JW, Francis ME, and Booth RJ. Linguistic Inquiry and Word Count. Mahwah, NJ: Erlbaum Publishers, 2001.

[21] Ramírez Esparza N, Pennebaker JW, García FA, and Suriá Martínez R. La psicología del uso de las palabras: un programa de computadora que analiza textos en español. Revista Mexicana de Psicología, 2007; 24(1): 85-89.

[22] Watson D, Clark L, and Tellengen A. Development and validation of brief measures of positive and negative affect: The PANAS scales. Journal of Personality and Social Psychology, 1988; 54(6): 1063-1070.

[23] móviles.com. El comparador de telefonía líder en España, http://www.moviles.com/ (accessed 17 June 2014).

[24] Cruz FM., Troyano JA, Enriquez F, and Ortega J. Clasificación de documentos basada en la opinión: experimentos con un corpus de críticas de cine enespañol. Procesamiento del lenguaje Natural, 2008; (41):73-80.

[25] Gholap J. Performance Tuning Of J48 Algorithm For Prediction Of Soil Fertility . Journal of Computer Science and Information Technology, 2012; 2(8).

[26] Pearl J. Bayesian networks: a model of self-activated memory for evidential reasoning. In: Proceedings of the 7th Conference of the Cognitive Science Society. Irvine, 1985, pp. 329-334.

[27] Platt J. Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines. Microsof Research, 1998.

[28] Keerti SS, Shevade SK, Battacharyya C, and Murthy K. Improvements to Platt's SMO Algorithm for SVM Classifier Design. Neural Computation, 2001; 13(3): 637-649.

[29] Nahar J, Tickle K, Ali S, and Chen P. Computational intelligence for microarray data and biomedical image analysis for the early diagnosis of breast cancer. Expert Systems with Applications, 2012; 39: 12371-12377.

[30] Chen L, Qi L, and Wang F. Comparison of feature-level learning methods for mining online consumer reviews. Expert Systems with Applications, 2012; 9588-9601.

[31] Chung C, and Pennebaker JW. The Psychological Functions of Function Words. Social Communication, 2007; 343-359.

[32] Bhavsar H, and Amit G. A Comparative Study of Training Algorithms for Supervised Machine Learning. International Journal of Soft Computing and Engineering (IJSCE). 2012; 2(4): 2231-2307.

[33] Chen YW, and Lin C J. Combining SVMs with various feature selection strategies. In: Feature Extraction Foundations and Applications. Studies in Fuzziness and Soft Computing, 2006, pp. 315-324.

[34] Tausczik YR, and Pennebaker JW. The psychological meaning of words: LIWC and computerized text analysis methods. Journal of language and social psychology, 2010; 29(1): 24-54.

[35] Baccianella S, Esuli A, and Sebastiani F. Sentiwordnet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. In Proceedings of the Seventh Conference on International Language Resources and Evaluation European Language Resources Association. 2010, pp. 2200–2204.

[36] Cambria E, Schuller B, Liu B, Wang H., and Havasi C. Knowledge-Based Approaches to Concept-Level Sentiment Analysis. IEEE Intelligent Systems. 2013; 28(2): 12-14.

[37] Ahmad T, and Doja MN. Rule Based System For Enhancing Recall For Feature Mining From Short Sentences In Customer Review Documents. International Journal on Computer Science & Engineering, 2012; 4(6).

[38] Feldman R. Techniques and Applications for Sentiment Analysis. Communications of the ACM, 2013; 56(4): 82-89.