Journal of Universal Computer Science, vol. 22, no. 5 (2016), 691-708
submitted: 15/9/15, accepted: 29/4/16, appeared: 1/5/16 © *J.UCS*

Sentiment Classification of Spanish Reviews: An Approach based on Feature Selection and Machine Learning Methods

María del Pilar Salas-Zárate
(Departamento de Informática y Sistemas, Universidad de Murcia, Spain
mariapilar.salas@um.es)

Mario Andrés Paredes-Valverde
(Departamento de Informática y Sistemas, Universidad de Murcia, Spain
marioandres.paredes@um.es)

Jorge Limon-Romero
(Facultad de Ingeniería, Arquitectura y Diseño, Universidad Autónoma de Baja California
Ensenada, Mexico
jorge.limon@uabc.edu.mx)

Diego Tlapa
(Facultad de Ingeniería, Arquitectura y Diseño, Universidad Autónoma de Baja California
Ensenada, Mexico
diegotlapa@uabc.edu.mx)

Yolanda Baez-Lopez
(Facultad de Ingeniería, Arquitectura y Diseño, Universidad Autónoma de Baja California
Ensenada, Mexico
yolanda@uabc.edu.mx)

Abstract: Sentiment analysis aims to extract users' opinions from review documents. Nowadays, there are two main approaches for sentiment analysis: the semantic orientation and the machine learning. Sentiment analysis approaches based on Machine Learning (ML) methods work over a set of features extracted from the users' opinions. However, the high dimensionality of the feature vector reduces the effectiveness of this approach. In this sense, we propose a sentiment classification method based on feature selection mechanisms and ML methods. The present method uses a hybrid feature extraction method based on POS pattern and dependency parsing. The features obtained are enriched semantically through common-sense knowledge bases. Then, a feature selection method is applied to eliminate the noisy and irrelevant features. Finally, a set of classifiers is trained in order to classify unknown data. To prove the effectiveness of our approach, we have conducted an evaluation in the movies and technological products domains. Also, our proposal was compared with well-known methods and algorithms used on the sentiment classification field. Our proposal obtained encouraging results based on the F-measure metric, ranging from 0.786 to 0.898 for the aforementioned domains.

Keywords: Sentiment Analysis, Opinion mining, Natural language processing, Machine learning, Feature selection methods.
Categories: I.2.7, I.2.2, I.7, H.3.3, L.3.2

1 Introduction

Sentiment analysis or opinion mining has become a popular topic since it enables the study of unstructured Web data in order to understand public opinion. In this regard, sentiment analysis is employed to extract users' reviews from textual data. The capture of public opinion is gaining momentum, particularly in terms of product preferences, marketing campaigns, political movements, financial aspects and company strategies. Through opinions, the users can know different point of view about a specific topic, and then take a decision based on such information. In order to create an automated system that performs an effective sentiment analysis, several authors have based their works on two main approaches: Semantic Orientation and Machine learning [Peñalver-Martinez et al., 2014]. The Semantic Orientation (SO) approach makes use of lexicons, such as WordNet-Affect [Valitutti, 2004] and SentiWordNet [Baccianella et al., 2010]. The Machine learning methods often rely on supervised classification approaches which require a set of features for the subsequent training of machine learning algorithm. The N-grams approach is commonly used to extract features [Arafat, et al., 2014] [Moraes et al., 2013], however, it provides a very high dimensionality of the feature space, which reduces the effectiveness of this approach. The Sentiment analysis based on Machine learning algorithms can address this issue by using feature selection methods that eliminate the noisy and irrelevant features, thus improving the classification accuracy [Gelbukh, 2013]. Among the feature selection methods commonly used in a supervised approach are Information Gain (IG) and minimum redundancy maximum relevance (mMRM) [Arafat et al., 2014] [Moraes et al., 2013] [Habernal et al., 2014].

In this work, we present a method for sentiment classification of Spanish reviews based on feature selection mechanisms and ML methods. The proposed method uses a hybrid feature extraction method based on POS pattern and dependency parsing. With this aim in mind, we use FreeLing [Padró et al., 2010], which is a multilingual system for linguistic analysis of texts, like tagging, lemmatization, and dependency parsing, among others. Also, the features obtained are enriched semantically through common-sense knowledge bases. Despite the fact that nowadays, there are several well-known common-sense knowledge bases such as WordNet [Miller, 1995] and Cyc [Lenat, 1995], in this work, we have opted for ConceptNet [Liu and Singh, 2004]. This decision is founded on two main reasons: (1) ConceptNet is a unique resource that captures a wide range of common-sense concepts and relations, such as those found in the Cyc knowledge base; (2) ConceptNet provides a more diverse relational ontology and emphasizes on informal conceptual connectedness over formal linguistic rigor. These features allow making practical, context-oriented, common-sense inferences over real-world texts [Speer and Havasi, 2013] [Speer and Havasi, 2012]. Regarding feature selection, the present method employs the Rough set theory (RST) and Information Gain (IG) algorithms to eliminate the noisy and irrelevant features, thus reducing the size of the feature vector previously obtained. Finally, these features were used for the subsequent training of the machine learning classifiers. In this work, three different classification algorithms were used, namely the Bayes Network learning algorithm (BayesNet), Maximum entropy(MaxEnt) and the SMO algorithm for SVM classifiers.

It is important to mention that most of the studies on polarity classification only deal with English documents, perhaps due to the lack of resources in other languages. Despite the fact that Spanish is currently the third language most used on the Internet according to the Internet World State rank[1], there are very few resources for managing sentiments or opinions in this language. Consequently, the management and study of subjectivity and sentiment analysis in languages other than English is a growing need. For this reason, this work is mainly motivated in the Spanish sentiment classification.

The remainder of the paper is structured as follows. Section 2 presents a review of the literature about machine learning approach for sentiment classification. The overall design of the proposed approach is described in section 3. Section 4 presents the evaluation results concerning the effectiveness of our approach to detect the polarity of Spanish reviews. Also, this section presents a comparison of our approach with well-known approaches used in the sentiment classification field. Finally, conclusions and future work are presented.

2 Related Work

In recent years, several researchers have introduced methods for sentiment classification. Most of these efforts are based on two approaches: the semantic orientation approach and the machine learning approach. It is worth mentioning that the present work strictly limits its scope to the latter approach. Therefore, in this section we will focus on the most related works in this area.

The machine learning approach often relies on supervised classification methods. These methods use a collection of data to train the classifier algorithms. Among the supervised machine learning techniques commonly used in the sentiment polarity classification are Support Vector Machine (SVM), Artificial Neural Network (ANN), Naive Bayes (NB), and Maximum Entropy (MaxEnt), among others. For example, [Moraes et al., 2013] presented an empirical comparison between SVM and ANN regarding document-level sentiment classification. Their experiments indicated that ANN produces superior results, especially on a benchmark dataset of movie reviews. It should be mentioned that ANN outperformed SVM by a statistically significant difference, even on the context of unbalanced data. On the other hand, [Salas-Zárate et al., 2014] presented a set of experiments that aimed to measure the effectiveness of the J48, SVM and BayesNet classification techniques on the sentiment classification for the Spanish. This study used the combination of the psychological and linguistic features of LIWC. The experiments were performed on the context of movies and technological products. The results indicated that SVM provided better results than the BayesNet and J48 algorithms. [Rushdi Saleh et al., 2011] applied a supervised machine learning method in order to classify reviews. They used SVM on three datasets with different sizes and domains. The first one [Pang and Lee, 2004] concerns movie reviews; the second corpus [Taboada and Grieve, 2004] concerns several topics like computers, hotels or music; and the last corpus was generated by crawling opinions about digital cameras from the Amazon website. The authors confirmed that SVM is a promising tool to deal with sentiment classification.

[1] http://www.internetworldstats.com/stats7.htm

On the other hand, machine learning methods require good, representative features for delivering good performance. Therefore, their success relies on the effectiveness of the feature extraction process. Most of the existing research focuses on features extracted through methods such as single words, character N-grams, and word N-grams (like bigrams and trigrams), as well as their combination. In [Agarwal and Mittal, 2014] methods such as unigrams, bigrams, dependency features and bi-tagged were used to extract features. Furthermore, new composite feature sets were generated from the features extracted by the above-mentioned methods. Experimental results showed that composite features constituted by prominent features resulting from unigram and bi-tagged methods perform a better sentiment classification than other methods. [Habernal et al., 2014] presented in-depth research on supervised machine learning methods for sentiment analysis of Czech Social media. The authors created a large Facebook dataset constituted by 10,000 posts, each of which was hand-annotated. Then, they evaluated the n-gram, character n-gram, POS-related and emoticons methods. The results showed that the combination of methods (unigrams, bigrams, POS features, emoticons, character n-grams) outperformed the baseline (unigram feature without preprocessing) with an F-measure of 0.69.

Other proposals have introduced methods for feature selection, aiming to improve the sentiment classification performance through the elimination of the noisy and irrelevant features of the feature vector previously obtained. [(Selvi et al., 2015] presented a comparative study of six feature selection mechanisms applied to machine learning, which are Information Gain (IG), Gain Ratio (GR), CHI squared, Relief-F (RF), and Significance Attribute Evaluation (SAE). The results showed that the Naïve Bayes machine learning method works better for the GR and SAE (Significance Attribute Evaluation) feature selection mechanisms. (Arafat et al., 2014) applied the RST (RougSet Theory), IG and mRMR (Minimum Redundancy Maximum Relevancy) feature selection methods on different data sets. The results showed that the mRMR provides better results than the IG method. On the other hand, [Agarwal and Mittal, 2013] used the IG and mRMR feature selection methods to extract prominent features. Further, the authors investigated the effect of using different feature sets for sentiment classification by using machine learning methods. In this sense, they used four standard datasets concerning movies, books, DVD’s and electronics. The experimental results showed that the mRMR method represents a better alternative for sentiment classification.

Table 1 summarizes the most relevant pieces of research aiming to fully analyze and compare them with our proposal. For this comparison, four features have been used: 1) classifier, 2) feature extraction, 3) feature selection method, 4) DataSet and 5) best result.

On the basis of the analysis presented above, our approach has some relevant and unique aspects. First, in the present work, we propose a method which aims to improve the sentiment classification based on machine learning taking into account three outstanding facts: 1) a set of composite features based on POS patterns and dependency parsing, 2) the collection of related common-sense knowledge from ConceptNet, and 3) the implementation of a hybrid feature selection method based on the RST and IG methods. Second, in order to validate the effectiveness of our approach we performed a comparison with the unigrams, bigrams, and mRMR methods. This latter method provides encouraging results in the works presented above.

Author	Classifier	Feature extraction	Feature selection method	Dataset	Best result
[Agarwal and Mittal, 2014]	BMNB	Composite unigrams and bi-tagged feature	mRMR	Movie and product	91.8 (F-measure)
[Habernal et al., 2014]	MaxEnt, SVM	n-Gram, Character n-gram, POS-related features, Emoticons.	Mutual Information (MI), Chi Square (CHI), Odds Ratio (OR), Relevancy Score (RS).	Movie and product	78.50 (F-measure)
[Arafat et al., 2014]	SVM, NB (Naïve Bayes)	Unigrams	IG, mRMR, IG-RS	Movie and product	87.7 (F-measure)
[Moraes et al., 2013]	SVM, ANN (Artificial Neural Networks)	Unigrams	IG	Movies, GPS, Books and Cameras	90.3 (Accuracy)

Table 1: Related work

3 Our approach

The sentiment classification method here proposed is divided into four main steps: (1) feature extraction, (2) increasing the semantic feature space, (3) feature selection and (4) training of the classifier. The first step consists on extracting the features from the text by means of the POS patterns and dependency parsing methods. The second step refers to increase the semantic feature space with common-sense knowledge. This task is performed by means of ConceptNet, a semantic network containing lots of things computers should know about the world, especially when understanding text written by people. The third step consists on the feature selection from the feature vector previously obtained. In this phase, the IG and RST feature selection methods are used. The last step consists of the training of the classifiers SVM, MaxEnt and BN. As will be shown in the following sections, the present approach has been tested through two Spanish corpora of different domains, especially movies and technological products. Figure 1 shows the flow diagram of our sentiment analysis approach.

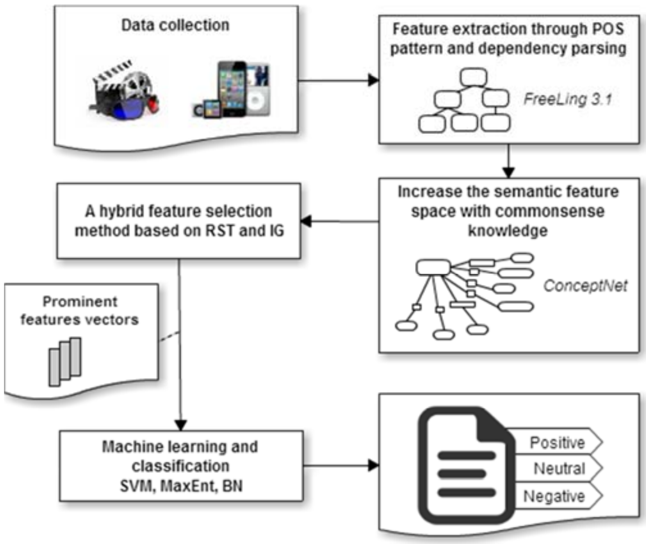

Figure 1: Flow diagram of our Sentiment Analysis approach

3.1 Feature extraction

The first phase of the method here presented consists on the extraction of a set of composite features by means of the combination of two extraction features methods, specifically the dependency parsing and POS pattern. The features extracted by the POS pattern method can include contextual information. However, this method is unable to extract syntactic information as in case of the dependency features method. This information is relevant in the sentiment analysis domain. Therefore, by combining the two above methods, it is possible to obtain more sentiment information in the form of contextual and syntactic patterns.

3.1.1 POS pattern based features

The first feature extraction method used in this work is the POS pattern. Aiming to perform this task, the POS-tagger for the Spanish language provided by FreeLing [Padró et al., 2010] is used. The POS-tagger analyzer allows to identify the lexical category of each word contained in the text. Some examples of these categories are the adjective, conjunction, adverb, verb, among others [Paredes-Valverde et al., 2015b]. The FreeLing software follows the EAGLES recommendations for morphosyntactic tag set. An excerpt of this recommendations is shown in Table 2. The full list of categories is presented in [Leech et al., 1996].

Tag	**Description**	**Example(s)**
AO0000	Adjective (ordinal)	primera, segundo, últimos
AQ0000	Adjective (descriptive)	populares, elegido, emocionada, andaluz
CC	Conjunction (coordinating)	y, o, pero
CS	Conjunction (subordinating)	que, como, mientras
RG	Adverb (general)	siempre, más, personalmente
RN	Adverb (negating)	no
VMM0000	Verb (main, imperative)	da, dé, trabaja, trabajes, trabajemos
VMN0000	Verb (main, infinitive)	dar, trabajar
VMP0000	Verb (main, participle)	dado, trabajado

Table 2: An excerpt of the EAGLES recommendations for morphosyntactic tag set.

We adopted the POS-based patterns approach presented in [Agarwal et al., 2015] aiming to extract two-word sentiment rich features from a sentence. For example, in the sentence "El teléfono es bueno y muy barato", whose POS-tagging processing is shown in Figure 2, the feature "muy barato" is obtained. However, as we can see in Figure 2, there is more sentiment-rich information which is intuitively very useful for the sentiment analysis. Such information can be extracted by using a dependency features approach. This approach is analyzed in the next section.

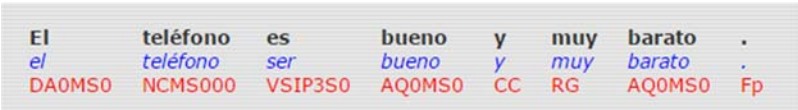

Figure 2: A POS-tagging example performed by Freeling.

3.1.2 Dependency features

As was previously mentioned, a deeper linguistic analysis aiming to extract syntactic relations contained in the natural language text is a very important task in the sentiment analysis process. In the literature, there are several works that have proved that the syntactic patterns approach is a very effective method for the subjective detection, which is a prior step to the sentiment classification. The present approach uses the Spanish FreeLing [Padró, 2012] dependency parser to extract all dependency relations found in the text. Figure 3 shows the dependency tree obtained by Freeling for the sentence "El teléfono es muy bueno y barato".

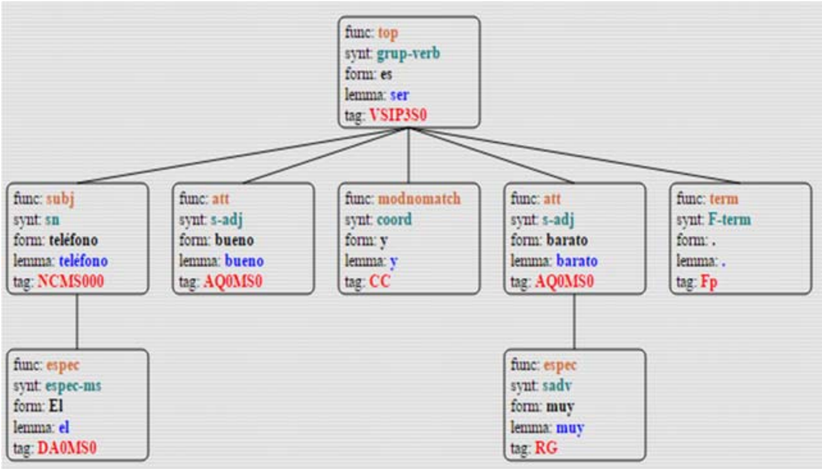

Figure 3: Dependency tree for the sentence "El teléfono es bueno y muy barato".

Aiming to extract the dependency features, we adopted the approach presented in [Agarwal et al., 2014]. This approach establishes a set of rules for the selection of sentiment-rich dependency features. Some of these rules are shown below.

- Copula. It is the relation between the complement of a copular verb and the copular verb. The Spanish language has two main copulative verbs, "ser" and "estar", which corresponds to the English verb "to be".
- Adverb modifier. The adverb modifier is used to essentially tweak the intensity of an adverb or adjective. This category includes Spanish adverbs such as "muy" (very), "más" (more) and "poco" (little). These adverbs generally come before the words they modify.
- Nominal subject. It is a noun phrase which is the syntactic subject of a clause. The governor of this relation might not always be a verb, when the verb is a copular verb, the root of the clause is the complement of the copular verb, which can be an adjective or a noun.

Based on the rules above, the relations copula(barato, es), copula(bueno, es), adverModifier(barato, muy), nominalSubject(barato, teléfono) and nominalSubject(bueno, teléfono) are extracted from the example show in Figure 3.

3.2 Collection of related common-sense knowledge

Common-sense knowledge is a collection of facts about the everyday words that are possessed by all people, i.e. it encompasses the spatial, physical, social, temporal, and psychological aspects of everyday life. Common-sense knowledge creation and usage are challenging fields because of the enormous breadth and details of common-sense knowledge. Many tasks like object recognition, machine translation and text mining require the machine to reach the human level of understanding to be done as well as a human being does it. This means that the machine should appear as intelligent as a human being [Alsoos and Kheirbek, 2015].

Nowadays, there are several well-known common-sense knowledge bases such as: WordNet [Miller, 1995], Cyc [Lenat, 1995], and ConceptNet [Liu and Singh, 2004]. WordNet is one of the most popular and widely used semantic resource in the computational linguistics and natural language processing community. It is a database of words that are categorized as nouns, verbs, adjectives and adverbs [Rodríguez-García et al., 2014]. WordNet contains only lexical structural relations such as synonym relation, which makes it semantically poor. On the other hand, Cyc tries to formalize common-sense knowledge in an upper knowledge base for free as a carefully designed ontology named OpenCyc. OpenCyc can be downloaded in OWL format or directly be accessed as an RDF store using web services from Cycorp. To use the Cyc reasoning engine, it is necessary to map the knowledge to its proprietary logical representation using the CycL language, which is a quite complex process. The difficulty of this mapping and the present unavailability of the full Cyc knowledge base to the general public, make Cyc an avoided choice in most cases. Finally, ConceptNet is a freely available multilingual common-sense knowledge base, which provides a large semantic graph that describes the general human knowledge and how it is expressed in natural language. ConceptNet defines concepts, which are words or phrases that can be extracted from natural language text. ConceptNet uses the term “concept” instead of “term” because of the fact that the words of phrases extracted can be more or less specific than a typical term. ConceptNet also contains assertions of the ways that these concepts relate to each other. Some of the most popular assertions are IsA, PartOf, MemberOf, RelatedTo, HasA, UsedFor, CapableOf, Synonym, Antonym and TranslationOf. These assertions can come from a wide variety of sources. Some current sources of knowledge in ConceptNet are The Open Mind Common Sense website, WordNet, and DBPedia [Speer and Havasi, 2013]. A partial snapshot of ConceptNet actual knowledge for the "good movie" concept is given in Figure 4.

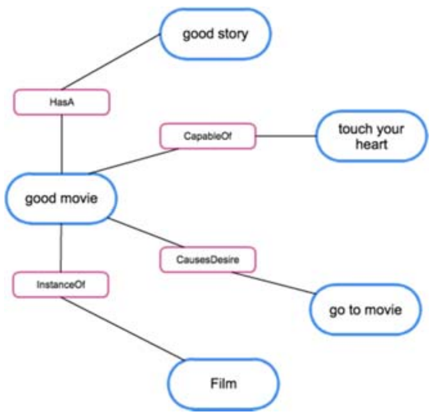

Figure 4: a partial snapshot of ConceptNet actual knowledge for “good movie” concept

As we can see in the above figure, the “good movie” concept can be extended through at least four semantic relations, for instance, we can establish that because it is a good movie, it also has a good story and it causes the desire of watching it.

The present approach uses ConceptNet in order to enrich semantically the feature vector previously obtained, i.e., this knowledge base is used to detect sentiment expressed explicitly through the analysis of features that not explicitly convey any sentiment but are explicitly linked to other feature that do convey sentiment.

3.3 Feature selection

The sentiment classification based on the machine learning approach faces the problem of high dimensionality of the feature vector. In this sense, a feature selection method is needed to eliminate the noisy and irrelevant features from the feature vector, thus improving the performance of machine learning algorithms. In this work, we use a hybrid feature selection approach based on the RST and IG feature selection methods. These methods are explained in detail below.

- The Rough Set Theory (RST) was developed by (Pawlak, 1982) at the Institute of Computer Sciences, in Warsaw. It is a mathematical approach for handling vagueness and uncertainty in data analysis. Objects may be indiscernible due to the limited available information. A rough set is characterized by a pair of precise concepts, called lower and upper approximations, which are generated using object indiscernibility. The most important issues are the reduction of attributes and the generation of decision rules. The rough set approach seems to be of fundamental importance to AI (Artificial Intelligence) and cognitive sciences, especially in the areas of machine learning, knowledge acquisition, decision analysis, knowledge discovery from databases, expert systems, inductive reasoning and pattern recognition [Arafat et al., 2014].
- Information Gain (IG) measures the relevance of a given feature. It is measured by the reduction in the uncertainty in identifying the class attribute when the value of the feature is known [Arafat et al., 2014]. The top ranked (important) features are selected for reducing the feature vector size aiming to obtain better classification results.

The hybrid feature selection method presented in this work is divided into two main phases. First, the IG of each feature is computed and all features with an IG value greater than 0 are selected. In this way, the irrelevant and noisy features are removed from the feature vector, therefore, the computational efforts for sentiment classification are reduced. The second phase is performed by means of the RST method, through which redundant information or features are removed. Then, a feature subset that has the same discernibility as the original set of features is selected.

3.4 Training of the classifiers

The final phase of the approach here proposed consists on the training of the classifiers. In this sense, we used WEKA, which is a collection of machine learning algorithms that can be used for classification and clustering. WEKA includes algorithms for classification, regression, clustering attribute selection and association rule mining. 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 present work uses three different classification algorithms, the Bayes Network learning algorithm (BayesNet), Maximum entropy(MaxEnt) and the SMO algorithm for SVM classifiers. These algorithms were selected because they have been used in several experiments obtaining good results in data classification [Xia et al., 2011], [He and Zhou, 2011], [Martín-Valdivia et al., 2012], [Montejo-Ráez et al., 2014]. The classifier training phase aims to build a model based on the analysis of the instances. This model is represented through classification rules, decision trees, or mathematical formulae. The models generated are used to classify unknown data. Aiming to measure the effectiveness of our approach to classify Spanish opinions in the movies and technological products domain we performed a set of experiments. These experiments are described in detail below.

4 Experiments and results

In this work, we evaluated our approach, aiming to measure its effectiveness to detect the polarity of Spanish opinions. In this evaluation process, two different datasets were considered (a) a dataset concerning movies domain and (b) a dataset concerning technological products. On the other hand, we performed a set of experiments in order to compare our approach with well-known approaches used on the sentiment classification field. The following sections describe in detail the aforementioned experiments.

4.1 Data

The set of experiments performed in this work involved the use of two datasets concerning the movie reviews and technological products domains. The first dataset was obtained from [Cruz Mata et al., 2008]. It 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 these experiments, we selected 300 random opinions of each category, making a total of 1,500 opinions. Figure 5 shows an example of a review with highly negative orientation.

```
<review author="Rafa Ferrer" title="Cuando llama un extraño"
rank="1" maxRank="5" source="muchocine">
<summary>
Hay que ser muy benévolo para calificar este engendro como
película, de lo peor del año 2006.
</summary>
<body>
Hay que ser muy benévolo para calificar este engendro como
película, sin pies, ni cabeza, con un guion realizado por un niño
de 4 años y una dirección que más que dar miedo, da risa, o miedo
de lo mala que es, según se mire. Es cierto que la cinta aporta una
pequeña dosis de tensión, creo que más que por la película, por la
obligación moral de disfrutar los 6 euro; de entrada. La
protagonista para hostiarla a mano abierta (deseé durante toda la
proyección que la mataran de una vez por todas) y el argumento una
especie de "refrito" de cintas como "Scream", "Hallowen" y "Viernes
13". Vaya, de lo peor que he visto en lo que llevamos de 2006.
</body>
</review>
```

Figure 5: Example of a review with highly negative orientation.

Regarding the second dataset used in these experiments, it contains 1500 reviews of technological products such as smartphones, laptops, tablets, among others, obtained from online selling websites such as [“moviles.com,” 2016]. This dataset is composed by 300 highly negative reviews, 300 negative reviews, 300 neutral reviews, 300 positive reviews and 300 highly positive reviews. It is worth noting that each review was examined and classified manually in order to ensure its quality. This time-consuming task was performed by a group of five people with a great experience in the sentiment classification domain. Figure 6 shows a review with highly positive orientation.

```
<review id=4 rank=”5” source=”moviles”>
<abstract>
Muy bueno
</abstract>
<content>
Tengo el galaxy s4 desde hace medio año, al principio me dio
problemas la batería,  investigué y encontré que mi problema era
algo llamado "muerte súbita" que consiste en que el teléfono se
apaga aun teniendo batería y no se enciende hasta que lo conectas
al cargador, lo llevé a la tienda samsung y en 5 min me cambiaron
la batería gratuitamente y desde entonces no he tenido problemas,
he de decir que hay aspectos del teléfono poco útiles como el poder
mover la pantalla con la mirada, pero este dispositivo es de lo
mejor que he visto. Personalmente no lo cambiaría por otro
teléfono, la cámara es genial. Soy diseñadora gráfica y por tanto
en fotografía sé lo que hablo y las fotos para ser hechas con un
móvil son geniales.
</content>
```

Figure 6: Example of review with highly positive orientation.

4.2 Evaluation and results

In order to measure the performance of our method, we have used three evaluation metrics that are commonly used in sentiment analysis: precision, recall and F-measure. Recall (1) is the proportion of actual positive cases that were correctly predicted as such. On the other hand, precision (2) represents the proportion of predicted positive cases that are real positives. Finally, F-measure (3) is the harmonic mean of precision and recall [Paredes-Valverde et al., 2015a].

$$(1)\ Recall = \frac{TP}{TP + FN}$$

$$(2)\ Precision = \frac{TP}{TP + FP}$$

$$(3)\ F1 = 2 * \frac{Precision * Recall}{Precision + Recall}$$

Table 3 presents the F-measure results obtained by our approach for the classification of technological products and movie reviews by using two, three and five categories: positive-negative, positive-neutral-negative, and highly positive-positive-neutral- negative-highly negative. As was previously mentioned, our approach uses three different classification algorithms, the BayesNet, MaxEnt and SMO algorithms. The results obtained by each algorithm are shown below.

	BN			**ME**			**SVM**		
	2	**3**	**5**	**2**	**3**	**5**	**2**	**3**	**5**
Technological products	0.842	0.821	0.786	0.859	0.841	0.813	0.882	0.871	0.85
Movies	0.875	0.854	0.819	0.885	0.867	0.839	0.898	0.887	0.866

Table 3: F-measure results obtained by our approach

As can be seen in the table above, the three classification algorithms used in these experiments obtained similar results, however SVM obtained the best results. The SVM algorithm has proved to be simpler, easier to implement, and generally faster. Also, these results can be justified by the analysis presented in [Bhavsar and Ganatra, 2012], where it is clearly shown how SVM models are more accurate in comparison to other classification algorithms such as: decision trees, neural network, Bayesian network, and nearest neighbor. The SVM algorithm has been successfully applied for text classification due to it is robust in high dimensional spaces and in scenarios where there is a sparse set of samples. Also, 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 [Deng et al., 2012].

Also, the classification results obtained show that with a smaller number of classes, the precision of our method increases, i.e. the classification with two classes provides better results than the classification with three and five classes. The above-mentioned situation is due to the fact that in a bipolar system there is less space for the classification of slippery cases. On the other hand, with regard to the corpora, our approach obtained better results through the movie reviews dataset. We ascribe this result to the fact that the technological product reviews dataset contains more comparative sentences than the movie reviews dataset, which makes this dataset be more difficult to classify.

4.3 Comparison with other methods

On the basis of the related work results presented in table 1, we can see that our proposal obtained similar results compared to these proposals. However, it is difficult to compare the different sentiment analysis 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 sentiment analysis methods would require the usage of the same testing corpus. For this reason, we carried out a comparison of our proposal with the methods implemented by these proposals. The Unigrams and Bigrams methods were used for the feature extraction process. These methods were selected due to they are the most widely used on supervised machine learning approaches for sentiment classification. Also, the IG and mRMR algorithms were used to the feature selection process. These algorithms were selected due to they have been considered among the best feature selection methods for sentiment classification.

The comparison experiments consisted of three phases. Firstly, the features were extracted from the above-presented datasets. This process was performed by using the Unigrams (Ng1) and Bigrams (Ng2) methods. The Unigrams method obtained the features by eliminating extra spaces and noisy characters between any two words. For example, in the sentence: “El móvil tiene excelentes características”. The features obtained are: “El”, “móvil”, “tiene”, “excelentes”, “características”. The Bigrams method obtained the features through the extraction of two consecutive words in the text. For example, in the sentence “Los actores son muy malos”. The bigram features obtained are: “Los_actores”, “actors_son”, “son_muy”, and “muy_malos”. Secondly, once the features were extracted, the IG and mRMR algorithms were used to select an optimal feature set. The IG algorithm selected the most important features regarding a class, whereas the mRMR algorithm selected features that have high dependency to class (maximum relevancy) and minimum dependency among features (minimum redundancy) [Arafat et al., 2014)]. Finally, the classifiers were trained based on the set of prominent features obtained by the previous phase. Table 4 shows the results obtained by our proposal and the above mentioned methods in the context of the technological products reviews.

	BN			**ME**			**SVM**		
	2	**3**	**5**	**2**	**3**	**5**	**2**	**3**	**5**
Ng1	0.704	0.683	0.648	0.732	0.714	0.686	0.772	0.761	0.74
Ng2	0.687	0.666	0.631	0.717	0.699	0.671	0.734	0.723	0.702
Ng1 & IG	0.711	0.69	0.655	0.739	0.721	0.693	0.779	0.768	0.747
Ng2 & IG	0.694	0.673	0.638	0.724	0.706	0.678	0.742	0.731	0.71
Ng1 & mRMR	0.719	0.698	0.663	0.747	0.729	0.701	0.787	0.776	0.755
Ng2 & mRMR	0.702	0.681	0.646	0.727	0.709	0.681	0.751	0.74	0.719
Our approach	0.842	0.821	0.786	0.859	0.841	0.813	0.882	0.871	0.85

Table 4: Evaluation results for the classification of technological products reviews

Table 5 shows the results obtained by our proposal and the above mentioned methods in the context of the movie reviews.

	BN			ME			SVM		
	2	3	5	2	3	5	2	3	5
Ng1	0.727	0.706	0.671	0.753	0.735	0.707	0.79	0.779	0.758
Ng2	0.71	0.689	0.654	0.738	0.72	0.692	0.752	0.741	0.72
Ng1 & IG	0.734	0.713	0.678	0.76	0.742	0.714	0.797	0.786	0.765
Ng2 & IG	0.717	0.696	0.661	0.745	0.727	0.699	0.76	0.749	0.728
Ng1 & mRMR	0.742	0.721	0.686	0.768	0.75	0.722	0.805	0.794	0.773
Ng2 & mRMR	0.725	0.704	0.669	0.748	0.73	0.702	0.769	0.758	0.737
Our approach	0.875	0.854	0.819	0.885	0.867	0.839	0.898	0.887	0.866

Table 5: Evaluation results for the classification of movies reviews

As can be seen in Table 4 and Table 5, our approach obtained encouraging results, with an F-measure score of 88.2 % for the technological products dataset and 89.8 % for the movies reviews dataset. Also, the results show that our approach outperformed to the other widely used methods for sentiment classification. We ascribe this to the following: 1) The feature extraction method based on POS pattern and dependency parsing obtains more sentiment-rich information, which is intuitively very useful for the sentiment analysis. 2) The use of common-sense knowledge allows enrich semantically the feature vector. 3) The hybrid feature selection method based on RST and IG provides better performance than the IG and mRMR individually [Arafat et al., 2014)]. Despite that IG can only compute the importance of the feature, the RST method reduces most of the irrelevant and noisy features as well as the redundancy among the features. Also, the RST method has the advantage of considering the dependency of combination of features on decision attribute in contrast to other conventional feature selection methods.

5 Conclusions and Future Work

The contribution of this work was twofold. First, we provide a novelty sentiment classification method based on feature selection and ML methods. Our approach combines the use of two well-known algorithms (RST and IG) for the feature selection process aiming to eliminate noisy and redundant features, thus reducing the high dimensionality of the feature vector obtained by previous phases. As was mentioned, the high dimensionality of feature vector reduces the effectiveness of the sentiment analysis based on ML methods, therefore, our approach tries to deal with this problem. Second, we presented a comparison of our approach with different methods and algorithms that have proved to be successful tools in the sentiment classification field.

Despite the results obtained by our approach seem encouraging, we are aware that the method here proposed may be further extended with capabilities that allow to improve the performance of the feature extraction and selection processes, which in turn, improves the general performance of our approach. On the one hand, as we can remember, the feature extraction method used in our approach is based on a set of dependency parsing patterns adapted from approaches focused on the English language. In this sense, we are convinced that this dependency patterns set can be extended aiming to deal with the higher grammatical complexity of the Spanish language. This task would require the analysis of a bigger set of datasets from different domains. Also, regarding feature selection process, we need to perform a deeper comparison of the algorithms used in this work with others such as Mutual Information (MI), Information Gain Ratio (GR) and Chi Square (CHI), which have been used in other domains getting good results. On the other hand, we are considering to evaluate our approach by using new corpora concerning different domains such as finances and tourism. Finally, we are considering to carry out a comparison with semantic orientation approach, with the aim to find advantages or disadvantages with respect to the machine learning approach, and especially with the approach presented in this work.

Acknowledgements

María del Pilar Salas-Zárate and Mario Andrés Paredes-Valverde are supported by the National Council of Science and Technology (CONACYT), the Public Education Secretary (SEP) and the Mexican government.

References

[Agarwal and Mittal, 2013] Agarwal, B., & Mittal, N. (2013). Sentiment Classification using Rough Set based Hybrid Feature Selection. In In Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis (pp. 115–119).

[Agarwal and Mittal, 2014] Agarwal, B., & Mittal, N. (2014). Prominent feature extraction for review analysis: an empirical study. Journal of Experimental & Theoretical Artificial Intelligence, 0(0), 1–14. http://doi.org/10.1080/0952813X.2014.977830

[Agarwal et al., 2014] Agarwal, B., Mittal, N., & Sharma, V. (2014). Semantic Orientation-Based Approaches for Sentiment Analysis. In Case Studies in Intelligent Computing (Vols. 1–0, pp. 61–78). Auerbach Publications. Retrieved from http://www.crcnetbase.com/doi/abs/10.1201/b17333-5

[Agarwal et al., 2015] Agarwal, B., Poria, S., Mittal, N., Gelbukh, A., & Hussain, A. (2015). Concept-Level Sentiment Analysis with Dependency-Based Semantic Parsing: A Novel Approach. Cognitive Computation, 7(4), 487–499. http://doi.org/10.1007/s12559-014-9316-6

[Alsoos and Kheirbek, 2015] Alsoos, M., & Kheirbek, A. (2015). A Semantic Approach to Enhance HITS Algorithm for Extracting Associated Concepts using ConceptNet. Journal of Digital Information Management, 13(1), 55–62.

[Arafat et al., 2014] Arafat, H., Elawady, R. M., Barakat, S., & Elrashidy, N. M. (2014). Different feature selection for sentiment classification, 3(1), 137–150.

[Baccianella et al., 2010] Baccianella, S., Esuli, A., & Sebastiani, F. (2010). SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. Presented at the Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC'10), European Language Resources Association (ELRA). Retrieved from http://www.lrec-conf.org/proceedings/lrec2010/pdf/769_Paper.pdf

[Bhavsar and Ganatra, 2012] Bhavsar, H., & Ganatra, A. (2012). A Comparative Study of Training Algorithms for Supervised Machine Learning, 2(4), 74–81.

[Cruz Mata et al., 2008] Cruz Mata, F., Troyano Jiménez, J. A., Enríquez de Salamanca Ros, F., & Ortega Rodríguez, F. J. (2008). Clasificación de documentos basada en la opinión: experimentos con un corpus de críticas de cine en español. Retrieved from http://rua.ua.es/dspace/handle/10045/8067

[Deng et al., 2012] Deng, N., Tian, Y., & Zhang, C. (2012). Support Vector Machines: Optimization Based Theory, Algorithms, and Extensions. CRC Press.

[Gelbukh, 2013] Gelbukh, A. (2013). Computational Linguistics and Intelligent Text Processing: 14th International Conference, CICLing 2013, Karlovasi, Samos, Greece, March 24-30, 2013, Proceedings. Springer.

[Habernal et al., 2014] Habernal, I., Ptáček, T., & Steinberger, J. (2014). Supervised sentiment analysis in Czech social media. Information Processing & Management, 50(5), 693–707. http://doi.org/10.1016/j.ipm.2014.05.001

[He and Zhou, 2011] He, Y., & Zhou, D. (2011). Self-training from labeled features for sentiment analysis. Information Processing & Management, 47(4), 606–616. http://doi.org/10.1016/j.ipm.2010.11.003

[Leech et al., 1996] Leech, G., Barnett, R., & Kahrel, P. (1996, March 11). EAGLES. Recommendations for the morphosyntactic annotation of corpora.

[Lenat, 1995] Lenat, D. B. (1995). CYC: A Large-scale Investment in Knowledge Infrastructure. Commun. ACM, 38(11), 33–38. http://doi.org/10.1145/219717.219745

[Liu and Singh, 2004] Liu, H., & Singh, P. (2004). ConceptNet — A Practical Commonsense Reasoning Tool-Kit. BT Technology Journal, 22(4), 211–226. http://doi.org/10.1023/B:BTTJ.0000047600.45421.6d

[Martín-Valdivia et al., 2012] Martín-Valdivia, M.-T., Montejo-Ráez, A., Ureña-López, A., & Rushdi Saleh, M. (2012). Learning to Classify Neutral Examples from Positive and Negative Opinions, 18(16), 2319–2333.

[Miller, 1995] Miller, G. A. (1995). WordNet: A Lexical Database for English. Commun. ACM, 38(11), 39–41. http://doi.org/10.1145/219717.219748

[Montejo-Ráez et al., 2014] Montejo-Ráez, A., Martínez-Cámara, E., Martín-Valdivia, M. T., & Ureña-López, L. A. (2014). Ranked WordNet graph for Sentiment Polarity Classification in Twitter. Computer Speech & Language, 28(1), 93–107. http://doi.org/10.1016/j.csl.2013.04.001

[Moraes et al., 2013] Moraes, R., Valiati, J. F., & Gavião Neto, W. P. (2013). Document-level sentiment classification: An empirical comparison between SVM and ANN. Expert Systems with Applications, 40(2), 621–633. http://doi.org/10.1016/j.eswa.2012.07.059

[moviles, 2016] moviles.com. (2016). Retrieved from http://www.moviles.com/

[Padró, 2012] Padró, L. (2012). Analizadores Multilingües en FreeLing. Linguamática, 3(2), 13–20.

[Padró et al., 2010] Padró, L., Collado, M., Reese, S., Lloberes, M., Castellón, I., Supaero, I., & Sabatier, U. P. (2010). I.: FreeLing 2.1: Five Years of Open-Source Language Processing Tools. In Proceedings of 7th Language Resources and Evaluation Conference, La.

[Pang and Lee, 2004] Pang, B., & Lee, L. (2004). A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts. In Proceedings of the

42Nd Annual Meeting on Association for Computational Linguistics. Stroudsburg, PA, USA: Association for Computational Linguistics. http://doi.org/10.3115/1218955.1218990

[Paredes-Valverde et al., 2015a] Paredes-Valverde, M. A., Rodríguez-García, M. Á., Ruiz-Martínez, A., Valencia-García, R., & Alor-Hernández, G. (2015). ONLI: An ontology-based system for querying DBpedia using natural language paradigm. Expert Systems with Applications, 42(12), 5163–5176. http://doi.org/10.1016/j.eswa.2015.02.034

[Paredes-Valverde et al., 2015b] Paredes-Valverde, M. A., Valencia-García, R., Rodríguez-García, M. Á., Colomo-Palacios, R., & Alor-Hernández, G. (2015). A semantic-based approach for querying linked data using natural language. Journal of Information Science, 0165551515616311. http://doi.org/10.1177/0165551515616311

[Pawlak, 1982] Pawlak, Z. (1982). Rough sets. International Journal of Computer & Information Sciences, 11(5), 341–356. http://doi.org/10.1007/BF01001956

[Peñalver-Martinez et al., 2014] Peñalver-Martinez, I., Garcia-Sanchez, F., Valencia-Garcia, R., Rodríguez-García, M. Á., Moreno, V., Fraga, A., & Sánchez-Cervantes, J. L. (2014). Feature-based opinion mining through ontologies. Expert Systems with Applications, 41(13), 5995–6008. http://doi.org/10.1016/j.eswa.2014.03.022

[Rodríguez-García et al., 2014] Rodríguez-García, M. A, Valencia-García, R., García-Sánchez, F., & Samper-Zapater, JJ. (2014). Ontology-based annotation and retrieval of services in the cloud. Knowledge-Based Systems, 56, 15–25. http://doi.org/10.1016/j.knosys.2013.10.006

[Rushdi Saleh et al., 2011] Rushdi Saleh, M., Martín-Valdivia, M. T., Montejo-Ráez, A., & Ureña-López, L. A. (2011). Experiments with SVM to classify opinions in different domains. Expert Systems with Applications, 38(12), 14799–14804.

[Salas-Zárate et al., 2014] Salas-Zárate, M. del P., López-López, E., Valencia-García, R., Aussenac-Gilles, N., Almela, Á., & Alor-Hernández, G. (2014). A study on LIWC categories for opinion mining in Spanish reviews. Journal of Information Science, 40(6), 749–760. http://doi.org/10.1177/0165551514547842

[Selvi et al., 2015] Selvi, C., Ahuja, C., & Sivasankar, E. (2015). A Comparative Study of Feature Selection and Machine Learning Methods for Sentiment Classification on Movie Data Set. In D. Mandal, R. Kar, S. Das, & B. K. Panigrahi (Eds.), Intelligent Computing and Applications (pp. 367–379). Springer India. Retrieved from http://link.springer.com/chapter/10.1007/978-81-322-2268-2_39

[Speer and Havasi, 2012] Speer, R., & Havasi, C. (2012). Representing General Relational Knowledge in ConceptNet 5.

[Speer and Havasi, 2013] Speer, R., & Havasi, C. (2013). ConceptNet 5: A Large Semantic Network for Relational Knowledge. In I. Gurevych & J. Kim (Eds.), The People's Web Meets NLP (pp. 161–176). Springer Berlin Heidelberg. Retrieved from http://link.springer.com/chapter/10.1007/978-3-642-35085-6_6

[Taboada and Grieve, 2004] Taboada, M., & Grieve, J. (2004). Analyzing appraisal automatically. In In Proceedings of the AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications (pp. 158–161).

[Valitutti, 2004] Valitutti, R. (2004). WordNet-Affect: an Affective Extension of WordNet. In In Proceedings of the 4th International Conference on Language Resources and Evaluation (pp. 1083–1086).

[Xia et al., 2011] Xia, R., Zong, C., & Li, S. (2011). Ensemble of feature sets and classification algorithms for sentiment classification. Information Sciences, 181(6), 1138–1152.