A Hybrid Approach for Sentiment Analysis Applied to Paper Reviews

Brian Keith
Department of Computer and Systems Engineering, Universidad Católica del Norte
Antofagasta, Chile
brian.keith@ucn.cl

Exequiel Fuentes
Department of Computer and Systems Engineering, Universidad Católica del Norte
Antofagasta, Chile
exequiel.fuentes@ucn.cl

Claudio Meneses
Department of Computer and Systems Engineering, Universidad Católica del Norte
Antofagasta, Chile
cmeneses@ucn.cl

ABSTRACT

This article discusses the problem of extracting sentiment and opinions about a collection of articles on scientific reviews conducted under an international conference on computing in Spanish language. The aim of this analysis is on the one hand to automatically determine the orientation of a review of an article and contrast this approach with the assessment made by the reviewer of the article. This would allow scientists to characterize and compare reviews crosswise, and more objectively support the overall assessment of a scientific article. A hybrid approach that combines an unsupervised machine learning algorithm with techniques from natural language processing is proposed to analyze reviews, and part-of-speech (POS) tagging to obtain the syntactic structure of a sentence. This syntactic structure, along with the use of dictionaries, allows to determine the semantic orientation of the review through a scoring algorithm. A set of experiments were conducted to evaluate the capability and performance of the proposed approaches relative to a baseline, using standard metrics, such as accuracy, precision, recall, and the F_1-score. The results show improvements in the case of binary, ternary and a 5-point scale classification in relation to classical machine learning algorithms such as SVM and NB, but they also present a challenge to improve the multiclass classification in this domain.

CCS CONCEPTS

•Computing methodologies →Supervised learning by classification; Machine learning algorithms; •Information systems →Data mining; *Sentiment analysis;*

KEYWORDS

Opinion mining, Sentiment Analysis, Naive Bayes, Support vector machines, Part-Of-Speech tagging, Paper review analysis

1 INTRODUCTION

Sentiment analysis includes a great amount of tasks such as sentiment extraction and classification, subjectivity detection, opinion summary, and opinion spam detection, among others. To do these tasks accurately, it is necessary to face several challenges, particularly the meaning formalization of an opinion. For this purpose, a series of formalisms and math representations to express opinions have been developed.

An application area where opinion mining techniques have not been applied yet is the reviewing process of scientific articles. In addition, the scientific paper reviewing process is the main quality control mechanism for most scientific communities. This involves reviewing each paper in order to provide suggestions to authors for correcting and improving a paper, whether they think it can be published or must be rejected [5]. As in the sentiment analysis in the industry, there is a suggestion to use opinion mining for analyzing the orientation of scientific paper reviews. This paper shows the application of sentiment analysis on a data set consisting of paper peer reviews. The domain of scientific paper reviews presents some major challenges, such as: (1) Usually classes are unbalanced, because there is a strong bias towards negative opinions; (2) Different reviews usually vary in terms of the number of assessments; (3) Normally, there is not a clear correlation between the number of positive and negative opinions with the final evaluation made by reviewers. All these issues make this domain a challenge for opinion mining and sentiment analysis purposes.

Specifically, anonymous reviews taken from an international conference have been used as a data set. This conference is an academic/business event of informatics and computer engineering. Authors submitted their papers through EasyChair. The papers could be written in Spanish, English or Portuguese. A double blind review scheme was used to prevent biases during the evaluation of the different articles. An international reviewing committee was in charge of the evaluation of each paper. The papers were distributed among the reviewers according to their affinity to the corresponding research area. The reviewers evaluated the submitted papers and provided their comments and evaluations in Spanish and in some cases in English.

This paper aims to present the implementation of sentiment analysis methods in the area of scientific paper reviews as a proof of concept for future applications. The used techniques include a Bayesian classifier (NB), a classifier built on the basis of support vector machines (SVM), an unsupervised classifier in the form of a scoring algorithm based on Part-Of-Speech tagging [21] and keyword matching, and finally a hybrid method using both the scoring algorithm and SVM.

2 RELATED WORK

2.1 Sentiment analysis

Sentiment classification can be traditionally done in two ways: supervised and unsupervised based on semantics. The success of these techniques depends mainly on the appropriate extraction of the set of characteristics used to detect sentiments. The most used supervised techniques are support vector machines (SVM) and naïve Bayes (NB) classifier [32]. Machine learning solutions involve building classifiers from a collection of documents, where each text

can be represented as a bag of words [28, 45]. Also, it is common to use some stemming techniques and stop word elimination. In general, classifiers with a good behavior in the domain where they are trained do not show the same behavior in another domain since they are highly dependent on training data used [1]. Most of the literature is dedicated to domain specific solutions, and while there is much work towards cross domain opinion mining most solutions are domain dependant [16]. This article focuses on the domain of scientific paper reviews.

Unsupervised semantics-based methods use dictionaries in which different types of words are classified according to their semantic orientation [44]. Unlike traditional machine learning methods, semantics-based unsupervised methods are more dependent on their domain, although their performance may vary from one domain to another. There are two important sub-categories to mention: dictionary-based and corpus-based. The dictionary-based technique uses a set of initial terms usually manually collected. This set grows by looking up synonyms and antonyms. An example of this type of dictionary is WordNet, which was used for developing SentiWordNet [2]. The main drawback of this type of approach is its inability to face the specific orientations of a domain and context. The corpus-based technique emerged with the purpose of providing dictionaries for a specific domain. These dictionaries result from a set of opinions seeds growing through the search of words related by means of statistical or semantic techniques such as Latent Semantic Analysis (LSA) or just by the frequency of occurrence of words within the collection of documents used [35].

Authors in [24] present a refined characterization of sentiment analysis techniques, including machine learning (supervised and unsupervised algorithms) and lexicon-based approaches (dictionary-based and corpus-based methods). In this review, supervised methods used for sentiment analysis include decision trees, support vector machines, neural networks, and methods based on probability, such as naive Bayes, Bayesian networks and maximum entropy.

A series of related papers is discussed below. Since there are no applications in the same domain, the domain of reviews or entity critique (e.g. films, hotels, products) is used as a reference since they are the closest among possible applications. This study is partially based on the work proposed by the authors in [47], where an opinion classification system of film reviews in Spanish is shown, using dependency parsing and POS tagging.

Table 1 shows results from different studies to determine polarity, starting with the seminal work from Pang et al. (2002). These results are shown with the purpose of providing a reference framework to evaluate results obtained. The table focuses mostly in binary classification. Not all the papers shown in the table will be discussed, unless they are pertinent to specific work.

The strategy used is shown in the Approach (App.) column. It can be based on machine learning (ML), lexicon (L) or it may be hybrid (H). The area being worked out is shown in the Domain column. Most work is done on film critiques or Twitter. The values in the Results column are shown in terms of general accuracy, unless otherwise stated. The best results obtained for a certain paper are shown. If work involves doing tests on different data sets or with different class amounts, results will be reported separated by a slash (/) in the same order. The information in the table was obtained from the systematic reviews in [32] and [39]. The first paper deals with opinion mining as a whole, while the second one focuses on deep learning, a machine learning branch with different applications in opinion mining.

Table 1: Results obtained from previous related works.

Year	App.	Domain and Authors	Result
2002	ML	Movie reviews. Pang et al. [28]	82.9%
2009	ML	Product reviews in English, Dutch and French. Boiy et al. [4]	83.30% / 69.80% / 67.68%
	L	Movie and product reviews translated to Spanish. Brooke et al. [6]	71.81%
2011	L	Movie reviews. Taboada et al. [38]	76.37%
	H	Twitter. Zhang et al. [49]	85.40%
2012	ML	Forum comments. Ortigosa-Hernández et al. [26]	83.63%
2013	ML	Movie reviews. Socher et al. [37]	85.40% / 45.70%
	H	Tourism product reviews. Marrese-Taylor [23]	85.50% / 75.50%
2014	H	Movie reviews. Poria et al. [31]	86.21%
	ML	Twitter. Tang et al. [42]	87.61% / 70.40%
2015	ML	Reviews in Japanese. Shi et al. [36]	89.30% F_1
	ML	Movie and product reviews. Tang et al. [41]	60.80% / 43.50%
	ML	Movie and product reviews. Tang et al. [40]	67.60% / 45.30%
	ML	Movie reviews. Li et al. [20]	86.50%
	L	Movie, hotel and product reviews. Vilares et al. [47]	78.50% / 80.11% / 89.38%
2016	H	Twitter. Ketan et al. [18]	63.23%
	ML	Movie and product reviews. Joulin et al. [17]	66.6% / 45.2%
	ML	Movie and product reviews. Yang et al. [48]	75.8% / 63.6%
	H	Twitter. Ghiassi et al. [12]	95.1%
	L	Movie reviews. Cambria [8]	90.1%

An effective sentiment analysis requires not only considering words individually, but also the linguistic construction of the sentence analyzed since it may totally change the sentiment expressed. The usual way of facing these constructions is by defining a heuristic. Authors in [28] work on film critiques and use a simple heuristic assuming that the negation scope includes words between the negator and the first punctuation after the negative term. Authors in [38] use data generated from the POS tagging process to identify the negation scope.

Apart from linguistic aspects, sentiment analysis must take into account the quality of the text analyzed. Furthermore, people make spelling and grammar mistakes. Some incorrectly written words were found during data processing. To solve these problems, spelling correctors may be used.

Research in the opinion mining area has greatly grown in the last decade, though most work focuses on texts written in English, for example, the paper proposed in [22] , where an opinion mining system is developed to identify preferences over tourist products in Los Lagos region, Chile. The study is interesting, but the set of

opinions used is written in English. These opinions were taken from the website TripAdvisor. While sentiment analysis in Spanish does not differ in essence with respect to English sentiment analysis, there is a lack of tools and libraries in comparison with English, which makes the implementation of sentiment analysis in Spanish more complex in general. Additionally, the Spanish language is less structured, compact and technical than English, which makes its semantic analysis difficult. Furthermore, only a small percentage of the research work is based on the Spanish language, with the vast majority of them focused on the English language.

Lexicon and grammar differences between Spanish and English may have an impact on the performance of systems trained for a certain language. Categorizing an opinion as positive, negative or neutral seems a simple task; however, it is really complex, particularly when opinions are written in different languages. Authors in [4] have studied the impact of English, German, and French particularities.

Some opinion mining studies focus on the Spanish language. One of the most relevant is proposed in [7]. It uses a semantics-based model defining a collection of dictionaries to calculate sentiments. Another study recently proposed in [46] describes an opinion mining system that classifies the orientation of Spanish texts taken from Twitter, according to an analysis of natural language, obtaining the syntactic sentence structure.

While works of sentiment analysis centered in movie reviews and product reviews are common in the literature, it must be mentioned that these domains of application are quite different from scientific paper reviews. An important difference is that peer reviews of research articles are an occluded genre (i.e. the documents are not publicly available) [14], contrary to movie reviews and product reviews that are intended for the general public.

Another key difference is the vocabulary used, which due to the scientific background of the domain tends to formality. An important difference is that in terms of the use of language the reviewers tend to respect the respective rules of ortography and grammar, which facilitates the analysis in comparison with the other kind of reviews. In general, the main difference is the expected level of formality found throughout the text.

Furthermore, the interpretation of a paper review can be a difficult task because of the conflicting signals contained in the text [13]. Also, reviews contain requests for changes in the form of directions, suggestions, clarification requests and recommendations. Early career researchers tend to be more affected by this, since they lack the experience to adequately interpret the reviewers' comments [27].

Finally, it is important to remark that no publications using scientific paper reviews as a work domain for sentiment analysis have been found in the related papers revised. So, this proposal for applying sentiment analysis is a novel contribution to this domain.

2.2 Applications

One of the common problems in scientific paper reviews is that the scores provided by reviewers can be inconsistent with what is written in the review. Particularly, there are cases in which reviewers are too strict, leading to the contradiction that, in reading the review, critiques are scarce, thus indicating that a paper was accepted, but in reading the reviewer's result, you may find that it was rejected. The problem can also be the opposite, that is, a reviewer makes substantive critiques while indicating that the paper must be accepted.

Concerning the problem above, consistency evaluation between the written review and the reviewers' score is proposed as a practical application of sentiment classification. For these reasons, the classifier used in this study was trained according to manual data tagging, not the reviewer's original classification. This allows revising the consistency between what the review states and what the reviewer says about the paper acceptance or rejection.

In this context, conducting a longitudinal evaluation of the consistency between the review and each reviewer's acceptance is proposed as future work. This evaluation must be done while keeping anonymity and giving each reviewer a numerical identifier so as to avoid revealing their true identity.

This work would allow classifying reviewers between strict (i.e., the score is always more negative than the review's critique) and non-strict (i.e., the score is always more positive than the review's critique). This classification can be applied in such a way that reviewers may be distributed equitably, thus guaranteeing that a good paper will not be rejected because reviewers are too strict and a poor paper will not be accepted because reviewers are not very strict.

The current system is used as a proof of concept, showing that it is possible to use automatic sentiment classification methods to determine review orientations. Certainly, the classification provided by the system is not expected to be consistent with the results given by the reviewers themselves. In fact, this is the consistency to be determined.

3 MATERIALS AND METHODS

3.1 Data set description

The data set consists of paper reviews sent to an international conference in Spanish [1]. It has a total of $N = 405$ instances evaluated with a 5-point scale, expressing the reviewer's opinion about the paper ("-2": very negative, "-1": negative, "0": neutral, "1": positive, "2": very positive). The attributes of each instance in the data set are described in Table 2.

Empty reviews and reviews in English are not considered in the analysis. Table 3 shows a basic statistics summary concerning word count and number of sentences for the reviews in the data set.

Figure 1 shows the data distribution in terms of the classifications assigned by the authors when reviewing the content of each review, note that the data set is skewed. Figure 2 shows the data distribution in terms of the classifications assigned by original reviewers. The distribution of the original scores is more uniform in comparison to the revised scores. This difference is assumed to come from a discrepancy between the way the paper is evaluated and the way the review is written by the original reviewer.

The study focuses on classifying reviews according to the scale determined by the authors. Original evaluations will be used as complements for evaluating the consistency between the classification inferred from the text and the one assigned by the reviewer.

[1] The data set used in this study can be found in http://mii.ucn.cl/files/2814/8570/2080/reviews.json

Table 2: Attribute description for the paper reviews data set.

Attribute	Description
Timespan	A date associated with the year of conference, which in turn corresponds with the time the review was written. The data set includes four years of reviews worth of conferences.
Paper ID	This number identifies each individual paper from a given conference. The data set has 172 different papers.
Review ID	A serial number identifier for each review as a correlative with respect to each individual paper. (e.g. the second review of some paper would correspond to the number 2). The data set has a total of 405 reviews. Most papers have 2 reviews each.
Text	Comments and detailed review of the paper. This is read by the authors and the editing commission of the conference. The editors determine if the paper should be published or not depending on the reviews. There are 6 instances of empty reviews.
Remarks	Additional comments that can be read only by the editing commission of the conference. This is used in conjunction with the previous attribute to determine if the paper should be published. This is an optional attribute. Whenever it is possible it is concatenated at the end of the main body of the review.
Language	Language corresponding to the review (it may be English or Spanish). In this case the majority of the reviews are in Spanish, with only 17 instances of English reviews.
Orientation	Review classification defined by the authors of this study, according to the 5-point scale previously described, obtained through the authors' systematic judgement of each review. This attribute represents the subjective perception of each review (i.e. how negative or positive the review is perceived when someone reads it).
Evaluation	Review classification as defined by the reviewer, according to the 5-point scale previously described. This attribute represents the real evaluation given to the paper, as determined by the reviewers.

Table 3: Review data set statistics

Statistic	Number of Words	Number of Sentences
Min	3	1
Max	530	47
Avg	88.64	8.91
Std. Dev	69.76	7.54

3.2 Tools and resources

The following tools were used for developing an opinion classification system and doing sentiment analysis:

(1) Python programming language, version 2.7.
(2) Scikit-learn library, its classifier implementations and evaluation methods [29].
(3) *Stanford POS Tagger* library, particularly its model for processing text in Spanish [43]. This model uses the form proposed by the EAGLES group to tag words [19] in each sentence.
(4) SentiWordNet 3.0 lexical ontology, containing semantic orientations and synonym sets in English [2]. A Spanish-translated version available in [30] was used. Some words and their translation were added to the original set because it was not complete.
(5) Dictionaries specifying the semantics of the words. They were constructed by manually reviewing the data set and finding words that fit in each category. The following dictionaries were considered:
    - Positive words, e.g. "bueno" (good) and "innovador" (innovative).
    - Negative, e.g "malo" (bad/wrong), "deficiente" (deficient).
    - Adversative words, e.g. "pero" (but).
    - Amplifier words, e.g. "muy" (very).
    - Mitigators words, e.g. "menos" (less).
    - Suggestion words, e.g. "sugerir" (to suggest), "corregir" (to correct).
    - Negation words, e.g. "no", "nunca" (never).
(6) A list of compound expressions that must be fused together before processing the text (e.g. "sin embargo", which is taken to be a single token in the form "sin_embargo").

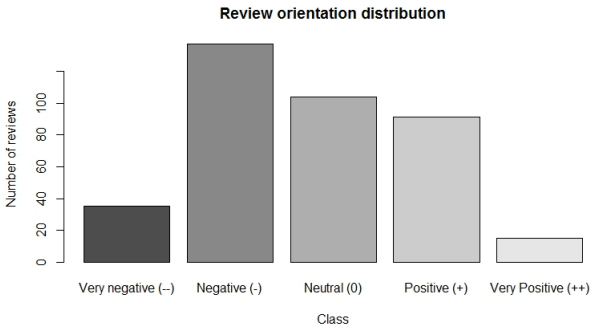

Figure 1: Distribution of review scores (revised value).

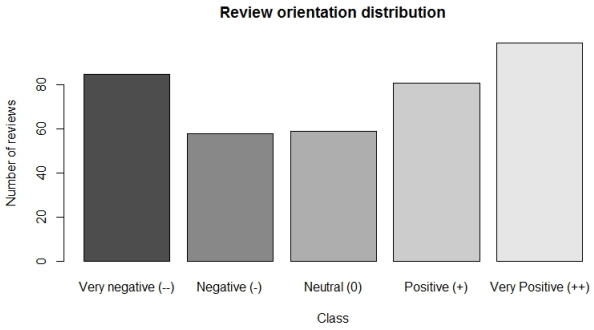

Figure 2: Distribution of review scores (original value).

3.3 Methods

Methods used in opinion mining are related to data extraction and preprocessing, natural language processing, and machine learning methods, which play a fundamental role in the task of determining the orientation of an opinion. A learning task may be divided into two broad approaches: supervised learning, in which classes are provided in data, and unsupervised learning, in which classes are unknown and the learning algorithm needs to automatically generate class values. Supervised methods naïve Bayes [3] and Support Vector Machines [21] were used. For the unsupervised learning task, an approach based on part-of-speech tagging and keyword matching was used. Furthermore, a hybrid approach [32] which combines both supervised and unsupervised methods is proposed.

Deep learning methods have not been tested due to the small size of the data set. While deep learning methods perform well in sentiment analysis [39], the number of parameters that must be estimated for deep learning to work well is too big for the amount of data present in this data set. Enlarging the data set is a difficult task since scientific reviews are an occluded genre [14] and as such getting access to more data is not easy. Gathering more reviews has been left for future work, and given this, the application of deep learning methods on this data set has been left for future work.

3.3.1 Supervised methods: NB and SVM. NB classifier assumes that all attributes are conditionally independent, but this assumption is not generally achieved in practice. For example, words in a document are not independent among them. Despite this, researchers have shown that this method generates good models [21].

As for SVM, this approach has a sound theoretical basis and has empirically shown to be the most accurate classifier for text documents [21]. The classifier implemented by Pyhton *scikit-learn* library based on [29], *libsvm* implementation [9] was used. Particularly, a linear kernel was used because it rendered better results than other nuclei available in the library (rbf, polynomial, and sigmoidal, see the test set). The optimal classifier parametrization was obtained via empirical tests. The optimal parameter C found corresponds to $C = 1.5$ (values from 0.5 to 3.0 with 0.25 increments were used). Default parameters were used for the other configurable elements of the implementation because they provided good results.

For SVM, an output coding based on error correction codes [10] was used. This method is implemented in *sklearn* libraries and its performance was better than the *one vs. all* approach used by default for the implementation [29], obtaining a 10% improvement in terms of the average metric F_1-score. The selected code size is twice greater than the amount of classes. This parameter was selected via empirical performance evaluation (values from 0.5 to 3.0, with 0.25 increments were tested).

In both cases, the training of the classifiers was done by splitting the data set into a training set and a testing set with a 70% and 30% proportion, respectively.

3.3.2 Unsupervised methods: Part-Of-Speech tagging. Once the text is separated in tokens, the next step is usually made to conduct a morphosyntactic analysis to identify characteristics, for example, its grammatical category. This analysis is known as Part-Of-Speech (POS) tagging.

The method uses a text in a given language as input and, through the application of its internal POS tagging model, assigns a grammatical category to the words in a sentence, for example, verb and adjective, among others. In addition, each category has its own characteristics, for example, in Spanish verbs are characterized by tense and type of subject, which are not applicable to nouns.

The complexity of this task depends on the target language to be analyzed. For example, Spanish is more complex as to verb conjugation and implicit subjects. To apply this technique, preprocessing stemming is omitted because it may prevent obtaining the correct grammar structure.

POS tagging poses two main challenges: The first one is word ambiguity, which depends on the context of the sentence analyzed; the second one is assigning a grammatical category to a word when the system does not know how to do it. To solve both problems, the context around the word in a sentence is typically considered and the most probable is selected. The grammatical category has a relevant characteristic. A word belonging to the same word group can replace a token with the same grammatical category, without affecting the sentence grammatically [33].

Most tools to determine grammatical category only work in English, as a result it becomes necessary to find a POS tagging library that can handle Spanish. The *Stanford Log-linear Part-Of-Speech Tagger* [15] library was used. This library reads a text and assigns a grammatical category to each word. This library is implemented in Java (version 8) and provides models in six different languages, including Spanish.

3.4 Data preprocessing

Before classifying a text, it is necessary to process it. First, punctuation standardization is done, so that writing rules can be respected (for example, "The writing is awful,but the form is correct." would become "The writing is awful, but the form is correct." (now, there is a space after the comma). Once this is done, the text is tokenized, separating it into sentences (according to the use of periods) and each sentence into words. Depending on each case, different preprocessing is done.

In the case of NB, punctuation marks and Spanish stopwords are eliminated because they do not provide any data for this classifier. A TF-IDF scheme is applied to the input text, this representation being Bayes classifier input. Similarly, in the case of SVM, punctuation marks and Spanish stopwords are eliminated. A TF-IDF scheme is applied to the input text; then, the singular value decomposition (SVD) method is applied, keeping 100 main values, this representation being SVM input. SVD is applied in order to reduce dimensionality, even though SVM is not sensitive to high dimensionalities, this reduction will reduce the computational cost of the method.

In the case of POS Tagging neither punctuation marks nor stopwords are eliminated because they contain useful data for the classifier (for example, negation). The text is then entered into Stanford POS Tagger in order to identify its semantic structure. Finally, a manual review is made to look for words (i.e. iterating over each word in the document) found in certain dictionaries so as to mark

these instances with additional tags. This list of tokens and their associated tags corresponds to the unsupervised classifier input.

3.5 Scoring Algorithm

To evaluate a review, Algorithm 1 is used over each sentence and then the average of all the sentences in the review are calculated. This average value provides the semantic orientation of the review in terms of a continuous numeric scale. This result must be discretized to obtain the classification in the corresponding classes.

The binary classification method (classes "-1" and "1"), ternary classification (classes "-1", "0", and "1"), and 5-point scale multiclass classification (from "-2" to "2") were tested, obtaining different performances in each case due to their increasing complexity.

The algorithm was implemented by following a rule-based scheme, according to the semantic characteristics of words. Particularly, a dictionary-based approach combined with a series of heuristics was used. Heuristics correspond to a series of rules that define the effect of each type of word on the semantic orientation of a sentence.

First, each word is analyzed to be tagged according to its semantic characteristics (POS Tagging). In addition, the dictionaries mentioned previously were used to add other tags in each word. The dictionaries are listed below, they were used in order to specify the effect of each word on the semantic orientation of the sentence. Particularly, the general effect on the sentence, according to a series of pre-established rules, is calculated, depending on the word found and its semantic orientation. The strategy used in each case is similar to the one used in [47], though without using dependency parsing.

(1) **Positive words:** It contains the list of words considered positive in the problem domain. Its semantic orientation is obtained from SentiWordNet ontology (values from 0 to 1), specifying that the positive value is required. In case the word is not in the ontology, a 0.5 default value is assumed (halfway between the minimum value of 0 and the maximum of 1, reflecting a moderately positive word).
(2) **Negative words:** It contains the list of words considered negative in the problem domain. Its semantic orientation is obtained from SentiWordNet ontology (values from 0 to -1), specifying that the negative value is required. In case the word is not in the ontology, a -0.5 default value is assumed (halfway between the minimum value of -1 and the maximum of 0, reflecting a moderately negative word).
(3) **Intensifiers:** It contains the list of words increasing the intensity of the words that follow by a certain predefined factor. The intensification factor is 2.5, a value that was considered empirically appropriate (values from 1.1 to 3.0 were tested, with 0.1 increments, the value 2.5 was chosen taking the value that provided the best average $F_1 - score$ based on a sample of 10 runs per value).
(4) **Mitigators:** It contains the list of words that decrease the intensity of the words that follow by a certain predefined factor. The reduction factor is 0.4 (values from 0.1 to 0.9 were tested, with 0.1 increments, the value 0.4 was chosen taking the value that provided the best average $F_1 - score$ based on a sample of 10 runs per value).
(5) **Negation words:** It contains the list of words that reverse the orientation of the words that follow (the semantic orientation value is multiplied by -1).
(6) **Adversative words:** It contains the list of adversative words. These reduce the intensity of previous words, but they intensify the ones that follow. There are two types of adversative clauses (restrictive and exclusive) [47]. While there exist other types of conjunctions (e.g. coordinate, copulative or disjunctive), for simplicity only adversative conjunctions were considered and for the purposes of this study, only the restrictive case was considered. The reduction factor is 0.5 (value empirically found; values from 0.1 to 0.9 were tested, with 0.1 increments).
(7) **Suggestion words:** It contains the list of words referring to a suggestion (for example, modal verbs like "should" and "could" and other verbs like "improve", "correct"). Modal verbs are very important due to the fact that they are emotional words giving either positive or negative polarity in the sentence. However, for this particular domain, these words are considered to have an always negative orientation that must be subtracted from the sentence score, however, they have a lesser impact in comparison to regular negative words.

    Usually, reviews that suggest direct rejection tend to use discourse units with the function of negative evaluation, while reviews that suggest a major revision of the article use discourse units with the function of recommendation [34]. Based on this, the score of a recommendation, while slightly negative in the sense that it implies that the paper must be improved, has a lower impact than a direct negative evaluation. The suitable empirical value was found to be -0.025 (value empirically found; it was tested from -0.5 to 0.0, with 0.025 increments).

In addition, four other heuristics not based on dictionaries were considered:

(1) If a question mark is found in the review, it causes a slight predefined negative impact, regardless of the context, which must be subtracted from the sentence score.
(2) If a negation adverb is found ("not") and it is followed by a verb, it has a strong predefined negative impact. The scope of the negation is considered to be up to three tokens after the adverb, based on the conservative heuristic used by Fernández Anta et al. [11]. To detect these patterns, POS tags are used.
(3) Bias parameters are included to strengthen positive and negative words. Since most reviews are likely to be critiques, it may be useful to include a bias towards positive opinion to compensate for the natural negativity. Movie reviews present similar behaviour, and bias parameters have been found to be useful [47].
(4) In case the word is not included in a dictionary (the list of words, not the ontology), it was assumed that it does not have any effect in this domain. So, its score is assigned to 0, under the assumption that it will have no effect.

Algorithm 1 Scoring Algorithm

Require: TokenList, a list of tokens in a sentence; PosBias, an additional weight factor for positive words; NegBias, an additional weight factor for negative words.

Ensure: TotalScore, the semantic orientation value for the sentence.

```
1:  function ScoreSentence
2:      TotalScore = 0
3:      PreviousTokens(2) = None
4:      Inverted = False
5:      TokenScore = 0
6:      for all (Token token in TokenList) do
7:          Tags = GetTags(Token)
8:          TokenScore = GetSentiWordNetScore(Token, Tags)
9:          if IsPositive(Tags) then
10:             TokenScore = TokenScore * PosBias
11:         else if IsNegative(Tags) then
12:             TokenScore = TokenScore * NegBias
13:         if Token == '?' then
14:             TokenScore = - QMOrientation
15:             Next Token
16:         if IsSuggestion(Tags) then
17:             TokenScore = - SuggestionOrientation
18:         if IsInversion(Tags) then
19:             Inverted = ¬ Inverted
20:         if Inverted then
21:             TokenScore = - TokenScore
22:         if IsVerb(Tags) and ContainsNo(PreviousTokens) then
23:             TotalScore = TotalScore - NegatedVerbOrientation
24:         if IsIncrement(PreviousTokens) then
25:             TokenScore = TokenScore * ModFactor
26:         if IsDecrement(PreviousTokens) then
27:             TokenScore = TokenScore / ModFactor
28:         if IsAdversative(Tags) then
29:             TotalScore = TotalScore * AdversativeWeight
30:         TotalScore = TotalScore + TokenScore
31:         Update PreviousTokens
    return TotalScore
```

These heuristics could be refined. Nevertheless, results obtained with them are satisfactory since the result improved compared to the baseline without using heuristics.

Algorithm 1 produces continuous values that can be positive or negative. Nevertheless, the objective is to obtain the semantic orientation in terms of the classes defined above. For this purpose, some parameter values (*DoublePositiveThreshold*, *DoubleNegativeThreshold*, *NegativeThreshold* y *PositiveThreshold*) were obtained by applying Monte Carlo simulation, testing a series of value ranges between -1.0 and 1.0 and using the combination with the best performance.

3.6 Hybrid method: HS-SVM

Another method based on the scoring algorithm and support vector machines is proposed for classification in this domain. The method has been named Hybrid Scoring Support Vector Machine (HS-SVM), in reference to the fact that it is a hybrid method that uses the scoring algorithm proposed in the previous section. This is a hybrid method of sentiment analysis since it combines a supervised classifier (SVM) and an unsupervised classifier (Scoring algorithm) to obtain the final class. The preprocessing steps for this new method are the same ones used for the original classifiers. Figure 3 shows the proposed method's components and flow.

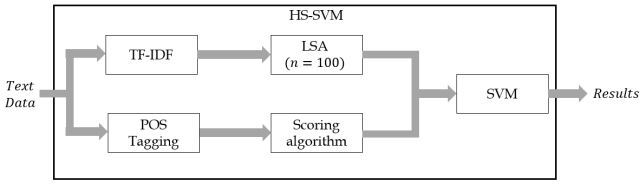

Figure 3: Hybrid Scoring Support Vector Machine components and flow.

The score works as a new feature for the SVM's input data. The SVM is then trained with this additional feature. This proposed approach has the advantage of having the information provided by the scoring algorithm and its associated components and the flexibility of the SVM. However, it has a higher computational cost since it requires the usage of the scoring algorithm and training the SVM classifier. Nevertheless, since the data set for this application is sufficiently small, this drawback has no significant effect.

3.7 Aspect Evaluator

Reviewer comments can have different functions, and they can be more directed towards the technical content, the general readibility or the structural aspect of the paper itself [14]. So while there are many aspects that could be evaluated, for example the opinion of the reviewer on the validity of the claims in the article or the discussion itself, it is simpler to evaluate textual aspects such as the format or writing rather than the content itself, since the latter requires certain knowledge of the domain of the reviewed article. Given this, a list of five important aspects considered when reviewing a paper was constructed. The evaluated aspects are References, Format, Structure, and Writing.

Evaluation consists in looking for references to these aspects (or their synonyms) in a sentence. A score is assigned to each sentence using Algorithm 1. The search of synonyms is done by using SentiWordNet synonym sets or synsets [2].

A vector containing the scores of each aspect is initialized in zero. As the algorithm evaluates the sentence tokens, POS tags are used to check if the current token is an adjective, a verb or a noun. These three tags were considered because an adjective and a verb may implicitly correspond to one aspect (e.g., "do not refer" or "well written"). If they correspond to one of these tags, they are checked to see if they agree with one of the aspects defined in the list. If all previous conditions apply, the current sentence score is added to the score of the associated aspect.

If an adversative clause is found, the current accumulated score is saved and a new accumulator is initialized because the use of these expressions marks the beginning of a different semantic orientation and the accumulation of previous values may affect the accuracy of results. The algorithm then continues its calculations using the

new accumulator. Once the algorithm finishes the analysis of the sentence, the final score is the sum of the old accumulator and the new accumulator.

In the final implementation, the scoring and aspect evaluation algorithms were considered as one function, for the sake of simplicity.

4 RESULTS AND DISCUSSION

4.1 Orientation classification

The results provided here originate from using the methods to classify the orientation of each review (i.e. the perceived evaluation). Table 4 shows the classification results for binary, ternary and 5-point scale classification.

Table 4: Classification results for orientation.

Binary				
Method	*Accuracy*	*Precision*	*Recall*	*F1*
NB	0.68 ± 0.05	0.67 ± 0.06	0.68 ± 0.05	0.64 ± 0.06
SVM	0.7 ± 0.05	0.7 ± 0.05	0.7 ± 0.05	0.69 ± 0.06
Score	0.81	0.81	0.81	0.81
HS-SVM	0.79 ± 0.05	0.8 ± 0.05	0.79 ± 0.05	0.79 ± 0.05
Ternary				
Method	*Accuracy*	*Precision*	*Recall*	*F1*
NB	0.46 ± 0.03	0.42 ± 0.05	0.46 ± 0.03	0.41 ± 0.05
SVM	0.48 ± 0.05	0.46 ± 0.06	0.48 ± 0.06	0.46 ± 0.06
Score	0.51	0.58	0.51	0.52
HS-SVM	0.56 ± 0.04	0.54 ± 0.04	0.56 ± 0.04	0.54 ± 0.04
5-point scale				
Method	*Accuracy*	*Precision*	*Recall*	*F1*
NB	0.35 ± 0.03	0.3 ± 0.04	0.35 ± 0.03	0.3 ± 0.04
SVM	0.4 ± 0.03	0.38 ± 0.04	0.41 ± 0.03	0.37 ± 0.03
Score	0.41	0.5	0.41	0.4
HS-SVM	0.46 ± 0.05	0.45 ± 0.06	0.46 ± 0.05	0.43 ± 0.05

In the binary case, performance is similar regarding the results from other studies (as shown in Table 1). The best average performance is obtained with the scoring algorithm, followed by HS-SVM, pure SVM and NB.

The amount of data available for the binary classification case is smaller than the amount of data for the multiclass case because the neutral reviews of the data set are not used. One of the main problems in comparison with other studies is the scarce amount of data available. A much better performance may be expected with a greater amount of instances.

In the case of ternary classification, average performance decreases in all metrics. This performance reduction is due to the greater classification complexity inherent to a problem with more classes. If the classifier were to work as a random selection it would only have 33.3% probability of predicting correctly. So, in comparison to that baseline, the classifiers still have a good quality. However, it is interesting to note that in this case, the best results are obtained with the HS-SVM classifier, which now surpasses the scoring algorithm itself.

The parametrization used is the one providing better results, according to the methodology explained in the previous section. There were problems with classifying the very negative reviews, if the lower limit is increased, examples of a very negative class can be correctly classified; however, some negative examples will also be incorrectly classified.

In the case of the 5-point scale classification, the scoring algorithm is slightly better than the supervised methods and the HS-SVM approach surpasses all the other methods in this case, just as it did in the ternary case. According to these results, the use of this hybrid approach has better classification performance in the multiclass case, while in the binary case it is only slightly behind the scoring algorithm. In this sense, this method is considered to be more robust in relation to an increase in the number of classes.

One of the main facts that may affect classification results for the supervised case is that these classifiers do not take into account text structure. They only consider the appearance of words according to the TF-IDF scheme described in the data preprocessing section.

The poor performance of SVM on this multiclass data set may be due to the fact that this classifier is highly sensitive to class imbalance [25]. And as Figure 1 shows, this data set is highly skewed. So, in a sense, the obtained results by SVM on that data set could not be reliable.

Better results could be obtained with the scoring algorithm by improving the heuristics used or applying parsing dependency [47]. Nevertheless, results are considered satisfactory, since in all the metrics this method surpasses the other approaches.

The performance improvement with respect to the pure SVM approach is consistent in all the cases. The method works by adding more information to SVM, basically facilitating the classification process. SVM is helped by the heuristics defined for the scoring algorithm.

This method could also be combined with the results obtained for the aspects of each review. In this approach, the use of the scoring algorithm and aspect evaluation could be considered as an additional preprocessing stage. This stage would have the function of calculating additional text characteristics to facilitate the classification process by supervised methods.

This combined approach may be used for generalizations in other opinion mining cases. It would be interesting to evaluate if similar improvements may be made in other domains. Certainly, it would be necessary to adapt and modify scoring algorithms and aspect evaluation, and probably obtain a new set of optimal parameters.

Adding a hierarchical classification approach may improve results, by first filtering neutral reviews, then applying binary classification, and later applying an approach on positive and negative sets to separate very negative/positive examples from those only negative/positive.

4.2 Evaluation classification

The results provided here are obtained from executing the methods to classify the evaluation of each review (i.e. the original score given by the reviewers). Table 5 shows the classification results for binary, ternary and 5-point scale classification.

In general, maximum possible performance decreases. Although the obtained results are still acceptable since they are better than a random selection, they show that properly classifying the instances is more complex if the original scores provided by each reviewer are used instead of the orientation scores. This discrepancy results

from the fact that reviewers do not usually provide scores agreeing with what is actually written in the review.

It is important to note that the parametrization of the scoring algorithm was not adjusted, retaining the original one designed for orientation classification. While this reduces classification accuracy and all associated metrics, this method is still competitive with the baseline methods (NB and SVM), and even those are still surpassed by the scoring algorithm classification in the binary case.

On the other hand, HS-SVM obtains the best results in comparison to the other methods. This stems from the flexibility provided by its SVM component, while at the same time benefiting from all the information provided by the scoring method. In general, according to the results of these experiments, HS-SVM surpasses the other methods, both in the evaluation classification task and in the orientation classification task.

Table 5: Classification results for evaluation.

Binary				
Method	*Accuracy*	*Precision*	*Recall*	*F1*
NB	0.56 ± 0.04	0.58 ± 0.04	0.56 ± 0.04	0.56 ± 0.04
SVM	0.67 ± 0.04	0.67 ± 0.04	0.67 ± 0.03	0.67 ± 0.04
Score	0.7	0.73	0.7	0.69
HS-SVM	0.71 ± 0.04	0.72 ± 0.04	0.71 ± 0.04	0.71 ± 0.04
Ternary				
Method	*Accuracy*	*Precision*	*Recall*	*F1*
NB	0.46 ± 0.04	0.45 ± 0.04	0.46 ± 0.04	0.44 ± 0.04
SVM	0.56 ± 0.04	0.53 ± 0.05	0.56 ± 0.04	0.53 ± 0.03
Score	0.46	0.62	0.46	0.5
HS-SVM	0.59 ± 0.02	0.56 ± 0.03	0.59 ± 0.02	0.57 ± 0.02
5-point scale				
Method	*Accuracy*	*Precision*	*Recall*	*F1*
NB	0.23 ± 0.02	0.27 ± 0.04	0.23 ± 0.02	0.24 ± 0.03
SVM	0.33 ± 0.05	0.35 ± 0.04	0.33 ± 0.05	0.33 ± 0.04
Score	0.27	0.55	0.27	0.24
HS-SVM	0.37 ± 0.06	0.38 ± 0.06	0.37 ± 0.06	0.36 ± 0.06

5 CONCLUSIONS

This article has studied the application of sentiment analysis techniques in the domain of paper reviews. Specifically, it has applied supervised methods (NB and SVM), an unsupervised method (the scoring algorithm) and a hybrid approach (HS-SVM) in the classification of 382 (non-empty Spanish) reviews of research papers of an international conference.

The best performance is obtained with binary classification, corresponding to the simplest version of the problem studied. Performance gradually decreases as more classes are added (such as the neutral one or those corresponding to extreme values). In this sense, the HS-SVM method is more robust than the others in relation to the number of classes.

One of the most interesting results is improvement obtained by the combination of the scoring algorithm and SVM. Basically, the score gives additional information to the SVM to facilitate the classification. Future work could deal with the extension and generalization of this method, also including the scores obtained for the aspects so as to further improve performance. By adding new semantic information (e.g. the score) to traditional machine learning methods, an improvement is expected to be obtained in the results of sentiment classification as compared with a pure method.

In the future, the algorithm performance to obtain the scores of each aspect must be evaluated. Its results were analyzed by observing those obtained in each review and the general average, but there is no specific metric as in the other methods evaluated. To better evaluate these results, it is necessary to have the tags for each aspect. These should be manually obtained in analyzing each review, although the weakness of this study is its subjectivity. So, automatic forms of generating tags for each aspect could be explored.

With respect to possible modifications of the models, one of the factors that could be considered in future work is individual reviewer bias (i.e. the reviewer may have a tendency to evaluate the papers lower or higher than the mean). In order to account for this bias, the current model would need to be modified. Also, another aspect that could be studied is an adequate handling of multi-lingual reviews, as well as the search of an appropriate parametrization in this case.

Concerning the experimental results, it is necessary to enlarge the list of features with more lexico-grammatical features, so that classifiers perform better and improved classification results are acquired. Also, expanding the data set with more reviews would be useful in future research, since the current data set is too small to apply some techniques that require more data to perform well.

As to the applicability of the proposal, future work could deal with the longitudinal evaluation of consistency between the review and the acceptance or rejection of the paper by each reviewer. This may allow a better evaluation of papers since it would be possible to recognize whether a reviewer is strict or not. Finally, since there are no other papers using scientific paper reviews as an application domain, the proposal in this study is a contribution and innovation for the field of sentiment analysis and opinion mining.

REFERENCES

[1] Anthony Aue and Michael Gamon. 2005. Customizing sentiment classifiers to new domains: A case study. In *Proceedings of recent advances in natural language processing (RANLP)*, Vol. 1. 2–1.

[2] Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining.. In *LREC*, Vol. 10. 2200–2204.

[3] Christopher Bishop. 2006. *Pattern Recognition and Machine Learning*. Springer-Verlag New York.

[4] Erik Boiy and Marie-Francine Moens. 2009. A Machine Learning Approach to Sentiment Analysis in Multilingual Web Texts. *Inf. Retr.* 12, 5 (Oct. 2009), 526–558. DOI : http://dx.doi.org/10.1007/s10791-008-9070-z

[5] Lutz Bornmann. 2011. Scientific peer review. *Annual Review of Information Science and Technology* 45, 1 (2011), 197–245.

[6] Julian Brooke, Milan Tofiloski, and Maite Taboada. 2009. Cross-Linguistic Sentiment Analysis: From English to Spanish.. In *RANLP*. 50–54.

[7] Julian Brooke, Milan Tofiloski, and Maite Taboada. 2009. Cross-Linguistic Sentiment Analysis: From English to Spanish.. In *RANLP*, Galia Angelova, Kalina Bontcheva, Ruslan Mitkov, Nicolas Nicolov, and Nikolai Nikolov (Eds.). RANLP 2009 Organising Committee / ACL, 50–54. http://dblp.uni-trier.de/db/conf/ranlp/ranlp2009.html#BrookeTT09

[8] Erik Cambria. 2016. Affective computing and sentiment analysis. *IEEE Intelligent Systems* 31, 2 (2016), 102–107.

[9] Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology* 2 (2011), 27:1–27:27. Issue 3. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[10] Thomas G. Dietterich and Ghulum Bakiri. 1995. Solving multiclass learning problems via error-correcting output codes. *Journal of artificial intelligence*

research (1995), 263–286.

[11] Antonio Fernández Anta, Philippe Morere, Luis F Chiroque, and Agustín Santos. 2012. Techniques for sentiment analysis and topic detection of Spanish tweets: preliminary report. (2012).

[12] Manoochehr Ghiassi, James Skinner, and David Zimbra. 2013. Twitter brand sentiment analysis: A hybrid system using n-gram analysis and dynamic artificial neural network. *Expert Systems with applications* 40, 16 (2013), 6266–6282.

[13] Hugh Gosden. 2001. Thank you for your critical comments and helpful suggestions: compliance and conflict in authors' replies to referees' comments in peer reviews of scientific research papers. *Ibérica: Revista de la Asociación Europea de Lenguas para Fines Específicos (AELFE)* 3 (2001), 3–17.

[14] Hugh Gosden. 2003. fiWhy not give us the full story?fi: functions of refereesfi comments in peer reviews of scientific research papers. *Journal of English for Academic Purposes* 2, 2 (2003), 87–101.

[15] The Stanford Natural Language Processing Group. 2015. Stanford Log-linear Part-Of-Speech Tagger. http://nlp.stanford.edu/software/tagger.shtml. (2015). http://nlp.stanford.edu/software/tagger.shtml Online; accessed 16-08-2015.

[16] Octavian Lucian Hasna, Florin Cristian Măcicăşan, Mihaela Dînşoreanu, and Rodica Potolea. 2014. Modeling Sentiment Polarity with Meta-features to Achieve Domain-Independence. In *International Joint Conference on Knowledge Discovery, Knowledge Engineering, and Knowledge Management*. Springer, 212–227.

[17] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759* (2016).

[18] Thakare Ketan Lalji and Sachin N Deshmukh. 2016. Twitter Sentiment Analysis Using Hybrid Approach. (2016).

[19] G Leech, R Barnett, and P Kahrel. 1996. Guidelines for the standardization of syntactic annotation of corpora. *EAGLES Document EAG-TCWG-SASG/1.8* (1996).

[20] Changliang Li, Bo Xu, Gaowei Wu, Saike He, Guanhua Tian, and Yujun Zhou. 2015. Parallel Recursive Deep Model for Sentiment Analysis. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 15–26.

[21] Bing Liu. 2011. *Web data mining: exploring hyperlinks, contents, and usage data*. Springer Science & Business Media.

[22] Edison Marrese. 2013. Diseño e implementación de una aplicación de web opinion mining para identificar preferencias de usuarios sobre productos turísticos de la X región de Los Lagos. http://www.repositorio.uchile.cl/handle/2250/113464. (2013). http://www.repositorio.uchile.cl/handle/2250/113464 Online; accessed 16-08-2015.

[23] Edison Marrese-Taylor, Juan D Velásquez, Felipe Bravo-Marquez, and Yutaka Matsuo. 2013. Identifying customer preferences about tourism products using an aspect-based opinion mining approach. *Procedia Computer Science* 22 (2013), 182–191.

[24] Walaa Medhat, Ahmed Hassan, and Hoda Korashy. 2014. Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal* 5, 4 (2014), 1093 – 1113. DOI: http://dx.doi.org/10.1016/j.asej.2014.04.011

[25] Asmaa Mountassir, Houda Benbrahim, and Ilham Berrada. 2012. An empirical study to address the problem of unbalanced data sets in sentiment classification. In *2012 IEEE international conference on systems, man, and cybernetics (SMC)*. IEEE, 3298–3303.

[26] Jonathan Ortigosa-Hernández, Juan Diego Rodríguez, Leandro Alzate, Manuel Lucania, Iñaki Inza, and Jose A Lozano. 2012. Approaching Sentiment Analysis by using semi-supervised learning of multi-dimensional classifiers. *Neurocomputing* 92 (2012), 98–115.

[27] Brian Paltridge. 2015. Referees' comments on submissions to peer-reviewed journals: when is a suggestion not a suggestion? *Studies in Higher Education* 40, 1 (2015), 106–122.

[28] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs Up?: Sentiment Classification Using Machine Learning Techniques. In *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10 (EMNLP '02)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 79–86. DOI: http://dx.doi.org/10.3115/1118693.1118704

[29] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[30] Esther Peinado. 2013. SentiWordNet-BD. https://github.com/rmaestre/Sentiwordnet-BC. (2013). https://github.com/rmaestre/Sentiwordnet-BC

[31] Soujanya Poria, Erik Cambria, Gregoire Winterstein, and Guang-Bin Huang. 2014. Sentic patterns: Dependency-based rules for concept-level sentiment analysis. *Knowledge-Based Systems* 69 (2014), 45–63.

[32] Kumar Ravi and Vadlamani Ravi. 2015. A survey on opinion mining and sentiment analysis: Tasks, approaches and applications. *Knowledge-Based Systems* 89 (2015), 14–46.

[33] M.J.F. Rodrigues and A.J. da Silva Teixeira. 2015. *Advanced Applications of Natural Language Processing for Performing Information Extraction*. Springer International Publishing. https://books.google.cl/books?id=PK0lCQAAQBAJ

[34] Betty Samraj. 2016. Discourse structure and variation in manuscript reviews: Implications for genre categorization. *English for Specific Purposes* 42 (2016), 76–88.

[35] Jesus Serrano-Guerrero, Jose A. Olivas, Francisco P. Romero, and Enrique Herrera-Viedma. 2015. Sentiment analysis: A review and comparative analysis of web services. *Information Sciences* 311 (2015), 18 – 38. DOI: http://dx.doi.org/10.1016/j.ins.2015.03.040

[36] Hanxiao Shi, Wenping Zhan, and Xiaojun Li. 2015. A Supervised Fine-Grained Sentiment Analysis System for Online Reviews. *Intelligent Automation & Soft Computing* ahead-of-print (2015), 1–17.

[37] Richard Socher, Alex Perelygin, Jean Y Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, Christopher Potts, and others. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the conference on empirical methods in natural language processing (EMNLP)*, Vol. 1631. Citeseer, 1642.

[38] Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. 2011. Lexicon-based Methods for Sentiment Analysis. *Comput. Linguist.* 37, 2 (June 2011), 267–307. http://dx.doi.org/10.1162/COLI_a_00049

[39] Duyu Tang, Bing Qin, and Ting Liu. 2015. Deep learning for sentiment analysis: successful approaches and future challenges. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 5, 6 (2015), 292–303.

[40] Duyu Tang, Bing Qin, and Ting Liu. 2015. Document Modeling with Gated Recurrent Neural Network for Sentiment Classification.. In *EMNLP*. 1422–1432.

[41] Duyu Tang, Bing Qin, and Ting Liu. 2015. Learning Semantic Representations of Users and Products for Document Level Sentiment Classification.. In *ACL (1)*. 1014–1023.

[42] Duyu Tang, Furu Wei, Bing Qin, Ting Liu, and Ming Zhou. 2014. Coooolll: A deep learning system for twitter sentiment classification. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*. 208–212.

[43] Kristina Toutanova, Dan Klein, Christopher D Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*. Association for Computational Linguistics, 173–180.

[44] Peter D. Turney. 2002. Thumbs Up or Thumbs Down?: Semantic Orientation Applied to Unsupervised Classification of Reviews. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL '02)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 417–424. DOI: http://dx.doi.org/10.3115/1073083.1073153

[45] Peter D Turney, Patrick Pantel, and others. 2010. From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research* 37, 1 (2010), 141–188.

[46] David Vilares, Miguel A. Alonso, and Carlos Gómez-Rodríguez. 2015. A Linguistic Approach for Determining the Topics of Spanish Twitter Messages. *J. Inf. Sci.* 41, 2 (April 2015), 127–145. DOI: http://dx.doi.org/10.1177/0165551514561652

[47] David Vilares, Miguel A Alonso, and Carlos Gomez-Rodriguez. 2015. A syntactic approach for opinion mining on Spanish reviews. *Natural Language Engineering* 21, 01 (2015), 139–163.

[48] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of NAACL-HLT*. 1480–1489.

[49] Ley Zhang, Riddhiman Ghosh, Mohamed Dekhil, Meichun Hsu, and Bing Liu. 2011. Combining lexicon-based and learning-based methods for twitter sentiment analysis. *HP Laboratories, Technical Report HPL-2011* 89 (2011).