A Hybrid Approach to Emotional Sentence Polarity and Intensity Classification

Jorge Carrillo de Albornoz, Laura Plaza, Pablo Gervás
Universidad Complutense de Madrid
Madrid, Spain
{jcalbornoz,lplazam}@fdi.ucm.es, pgervas@sip.ucm.es

Abstract

In this paper, the authors present a new approach to sentence level sentiment analysis. The aim is to determine whether a sentence expresses a positive, negative or neutral sentiment, as well as its intensity. The method performs WSD over the words in the sentence in order to work with concepts rather than terms, and makes use of the knowledge in an affective lexicon to label these concepts with emotional categories. It also deals with the effect of negations and quantifiers on polarity and intensity analysis. An extensive evaluation in two different domains is performed in order to determine how the method behaves in 2-classes (*positive* and *negative*), 3-classes (*positive*, *negative* and *neutral*) and 5-classes (*strongly negative*, *weakly negative*, *neutral*, *weakly positive* and *strongly positive*) classification tasks. The results obtained compare favorably with those achieved by other systems addressing similar evaluations.

1 Introduction

Sentiment analysis has gained much attention from the research community in recent years. It is concerned with the problem of discovering emotional meanings in text, and most common tasks usually include emotion labeling (assigning a text its main emotion), polarity recognition (classifying a statement into positive or negative) and subjectivity identification (determining whether a text is subjective or objective). The growing research interest is mainly due to the practical applications of sentiment analysis. Companies and organizations are interested in finding out costumer sentiments and opinions, while individuals are interested in others' opinions when purchasing a product or deciding whether or not watching a movie.

Many approaches have dealt with sentiment analysis as the problem of classifying product or service reviews (Pang et al., 2002; Turney, 2002), while others have attempted to classify news items (Devitt and Ahmad, 2007). The task is usually addressed as a 2-classes classification problem (*positive* vs. *negative*). Recent works have included the *neutral* class, trying to detect not only the polarity but also the absence of emotional meaning (Wilson et al., 2005; Esuli and Sebastiani, 2006). However, few approaches try to face a more fine-grained prediction of the intensity (e.g. classifying the polarity into *strongly negative*, *weakly negative*, *neutral*, *weakly positive* and *strongly positive*).

Another important problem of most of these approximations is that they usually work with terms, and so disregard the contextual meaning of those terms in the sentence (Martineau and Finin, 2009; Moilanen and Pulman, 2007). The use of word disambiguation is not usual in this task, due to the fact that most approaches use lexical resources created to work with terms. However, it is essential to correctly capture the meaning of these terms within the text.

In this paper, we present a hybrid approach based on machine learning techniques and lexical rules to classify sentences according to their polarity and intensity. Thus, given an input text, the method is able to determine the polarity of each sentence (i.e. if it is negative or positive), as well as its intensity. The system tackles the effect of negations and quantifiers in sentiment analysis, and addresses the problem of word ambiguity, taken into account the contextual meaning of the terms in the text by using a word sense disambiguation algorithm.

The paper is organized as follows. Section 2 exposes some background and related work on sentiment analysis. Section 3 presents the lexical resources and corpora used by the system. Sec-

Proceedings of the Fourteenth Conference on Computational Natural Language Learning, pages 153–161,
Uppsala, Sweden, 15-16 July 2010. ©2010 Association for Computational Linguistics

tion 4 describes the method proposed for polarity and intensity classification. Section 5 presents the evaluation framework and discusses the experimental results. Finally, section 6 provides concluding remarks and future lines of work.

2 Related work

The sentiment analysis discipline in computational linguistic is mainly focused on identifying/classifying different emotional contents within a phrase, sentence or document. This field usually encloses tasks such as emotion identification, subjectivity classification and polarity recognition. Sentiment analysis has obtained great popularity in the last years mostly due to its successful application to different business domains, such as the evaluation of products and services, where the goal is to discern whether the opinion expressed by a user about a product or service is favorable or unfavorable.

Focusing on polarity recognition, the aim of this task is the classification of texts into positive or negative according to their emotional meaning. Most of the approaches rely on machine learning techniques or rule based methods. Statistical approaches based on term frequencies and bags of words are frequently used in machine learning approximations. Pang et al. (2002) present a comparison between three different machine learning algorithms trained with bags of features computed over term frequencies, and conclude that SVM classifiers can be efficiently used in polarity identification. Martineau and Finin (2009) use a similar approach where the words are scored using a Delta TF-IDF function before classifying the documents. On the other hand, Meena and Prabhakar (2007) study the effect of conjuncts in polarity recognition using rule based methods over the syntax tree of the sentence. Whitelaw et al. (2005) introduce the concept of “appraisal groups” which are combined with bags of word features to automatically classify movie reviews. To this aim, they use a semi-automated method to generate a lexicon of appraising adjectives and modifiers.

During the past few years, the problem of polarity recognition has been usually faced as a step beyond the identification of the subjectivity or objectivity of texts (Wiebe et al., 1999). Different approximations have been proposed to deal with this problem. Pang and Lee (2004) propose a graph-based method which finds minimum cuts in a document graph to classify the sentences into subjective or objective. After that, they use a bag of words approximation to classify the subjective sentences into positive or negative. Kim and Hovy (2004) also introduce a previous step to identify the subjectivity of sentences regarding a certain topic, and later classify these sentences into positives or negatives.

Most recent approaches do not only deal with the 2-classes classification problem, but also introduce a new class representing neutrality. Thus, the aim of these works is to classify the text into positive, negative or neutral. Wilson et al. (2005) present a double subjectivity classifier based on features such as syntactic classes and sentence position, and more semantic features such as adjective graduation. The first classifier determines the subjectivity or neutrality of the phrases in the text, while the second determines its polarity (including neutrality). Esuli and Sebastiani (2006) also address this problem testing three different variants of a semi-supervised method, and classify the input into positive, negative or neutral. The method proposed yields good results in the 2-classes polarity classification, while the results decrease when dealing with 3-classes. A more ambitious classification task is proposed by Brooke (2009), where the goal is to measure the intensity of polarity. To this aim, the author classifies the input into 3-classes (*strongly-negative, ambivalent,* and *strongly-positive*), 4 classes (*strongly-negative, weakly-negative, weakly-positive* and *strongly-positive*) and 5-classes (*strongly-negative, weakly-negative, ambivalent, weakly-positive* and *strongly-positive*). The results decrease considerably with the number of classes, from 62% of accuracy for 3-classes to 38% of accuracy for 5-classes.

3 Corpora and resources

The evaluation of the system has been carried out using two corpora from two very distinct domains: the Sentence Polarity Movie Review Dataset[1] and the one used in the SemEval 2007 Affective Text task[2]. The first one consists of 10.662 sentences selected from different movie review websites. These sentences are labeled as positive or negative depending on whether they express a positive or negative opinion within the movie review. The second one consists of a training set and a test set of 250 and 1000 news headlines respectively, extracted from different news sites. Each sentence is labeled with a value

[1] http://www.cs.cornell.edu/People/pabo/movie-review-data/

[2] http://www.cse.unt.edu/~rada/affectivetext/

between -100 and 100, where -100 means highly negative emotional intensity, 100 means highly positive and 0 means neutral. To the purpose of this work, the test set from the SemEval corpus and 1000 sentences randomly extracted from the Sentence Polarity Movie Review corpus (500 positive and 500 negative) were used as evaluation datasets.

In order to identify the emotional categories associated to the concepts in the sentences, an affective lexical database based on semantic senses, instead of terms, is needed. To this aim, the authors have tested different resources and finally selected the WordNet Affect affective database (Strapparava and Valitutti, 2004). This affective lexicon has the particularity of assigning emotional categories to synsets of the WordNet lexical database (Miller et al., 1990), allowing the system to correctly disambiguate the terms using one of the many WordNet-based word sense disambiguation algorithms. The emotional categories in WordNet Affect are organized hierarchically, and its first level distinguishes between *positive-emotion, negative-emotion, neutral-emotion* and *ambiguous-emotion*. The second level encloses the emotional categories themselves, and consists of a set of 32 categories. For this work, a subset of 16 emotional categories from this level has been selected, since the hierarchy proposed in WordNet Affect is considerably broader than those commonly used in sentiment analysis. On the other hand, the first level of emotional categories may be useful to predict the polarity, but it is clearly not enough to predict the intensity of this polarity. To be precise, the subset of emotional categories used in this work is: *{joy, love, liking, calmness, positive-expectation, hope, fear, sadness, dislike, shame, compassion, despair, anxiety, surprise, ambiguous-agitation* and *ambiguous-expectation}*. The authors consider this subset to be a good representation of the human feeling.

Since the WordNet Affect hierarchy does not provide an antonym relationship, the authors has created that relation for the previous set of emotional categories. Only relationships between emotional categories with a strongly opposite meaning are created, such as *liking-disliking* and *joy-sadness*. The purpose of this antonym relationship is twofold: first, it contributes to handle negation forms; and second, it can be used to automatically expand the affective lexicon. Both issues are discussed in detail later in the document.

On the other hand, since a good amount of words with a highly emotional meaning, such as *dead*, *cancer* and *violent*, are not labeled in WordNet Affect, these words have been manually labeled by the authors and have been later extended with their *synonyms, antonyms* and *derived adjectives* using the corresponding semantic and lexical relations in WordNet. This process has been done in two steps in order to measure the effect of the number of synsets labeled on the classification accuracy, as described in section 5.

The WordNet Affect 1.1 lexicon consists of a set of 911 synsets. However, the authors have detected that a good number of these synsets have been labeled more than once, and with different emotional categories (e.g. the synset "a#00117872 {angered, enraged, furious, infuriated, maddened}" is labeled with three different categories: *anger, fury* and *infuriation*). Thus, after removing these synsets and those labeled with an emotional category not included in the 16-categories subset used in this work, the affective lexicon presents 798 synsets. After the first step of semi-automatic labeling, the affective lexicon increased the number of synsets in 372, of which 100 synsets were manually labeled, and 272 were automatically derived throughout the WordNet relations. The second and last step of semi-automatic labeling added 603 synsets to the lexicon, of which 200 synsets were manually labeled, and 403 were automatically derived. The final lexicon presents 1773 synsets and 4521 words labeled with an emotional category. Table 1 shows the distribution of the affective lexicon in grammatical categories.

Grammatical Category	**WNAffect**	**WNAffect + 1st step**	**WNAffect + 2nd step**
Nouns	280	440	699
Verbs	122	200	309
Adjectives	273	394	600
Adverbs	123	136	165

Table 1: Distribution in grammatical categories of the synsets in the affective lexicon.

4 The method

In this section, the method for automatically labeling sentences with an emotional intensity and polarity is presented. The problem is faced as a text classification task, which is accomplishes throughout four steps. Each step is explained in detail in the following subsections.

4.1 Pre-processing: POS tagging and concept identification

In order to determine the appropriate emotional category for each word in its context, a pre-processing step is accomplished to translate each term in the sentence to its adequate sense in WordNet. To this aim, the system analyzes the text, splits it into sentences and tags the tokens with their part of speech. The Gate architecture[3] and the Stanford Parser[4] were selected to carry out this process. In particular the *Annie English Tokeniser*, *Hash Gazetter*, *RegEx Sentence Splitter* and the *Stanford Parser* modules in Gate are used to analyze the input. In this step also the syntax tree and dependencies are retrieved from the Stanford Parser. These features will be used in the post-processing step in order to identify the negations and the quantifiers, as well as their scope.

Once the sentences have been split and tagged, the method maps each word of each sentence into its sense in WordNet according to its context. To this end, the *lesk* WSD algorithm implemented in the WordNet Sense-Relate perl package is used (Patwardhan et al., 2005). The disambiguation is carried out only over the words belonging to the grammatical categories *noun*, *verb*, *adjective* and *adverb,* as only these categories can present an emotional meaning. As a result, we get the stem and sense in WordNet of each word, and this information is used to retrieve its synset.

A good example of the importance of performing word disambiguation can be shown in the sentence "*Test to predict breast cancer relapse is approved*" from the SemEval news corpus. The noun *cancer* has five possible entries in WordNet and only one refers to a "*malignant growth or tumor"*, while the others are related with "*astrology"* and the "*cancer zodiacal constellation"*. Obviously, without a WSD algorithm, the wrong synset will be considered, and a wrong emotion will be assigned to the concept.

Besides, to enrich the emotion identification step, the hypernyms of each concept are also retrieved from WordNet.

4.2 Emotion identification

The aim of the emotion identification step is to map the WordNet concepts previously identified to those present in the affective lexicon, as well as to retrieve from this lexicon the corresponding emotional category of each concept.

We hypothesize that the hypernyms of a concept entail the same emotions than the concept itself, but the intensity of such emotions decreases as we move up the hierarchy (i.e. the more general the hypernym becomes, the less its emotional intensity is). Following this hypothesis, when no entry is found in the affective lexicon for a given concept, the emotional category associated to its nearest hypernym, if any, is used to label the concept. However, only a certain level of hypernymy is accepted, since an excessive generalization introduces some noise in the emotion identification. This parameter has been empirically set to *3* (Carrillo de Albornoz et al., 2010). Previous experiments have shown that, upper this level, the working hypothesis becomes unreliable.

The sentence "*Siesta cuts risk of heart disease death study finds*" clearly illustrates the process described above. In this sentence, the concepts *risk*, *death* and *disease* are labeled with an emotional category: in particular, the categories assigned to them are *fear, fear* and *dislike* respectively. However, while the two firsts are retrieved from the affective lexicon by their own synsets, the last one is labeled through its hypernym: since no matching is found for *disease* in the lexicon, the analysis over its hypernyms detects the category *dislike* assigned to the synset of its first hypernym, which contains words such as *illness* and *sickness*, and the same emotion *(dislike)* is assigned to *disease*.

It must be noted that, to perform this analysis, a previous mapping between 2.1 and 1.6 WordNet versions was needed, since the method and the affective lexicon work on different versions of the database.

4.3 Post-processing: Negation and quantifiers detection

Once the concepts of the sentence have been labeled with their emotional categories, the next step aims to detect and solve the effect of the negations and the quantifiers on the emotional categories identified in the previous step.

The effect of negation has been broadly studied in NLP (Morante and Daelemans, 2009) and sentiment analysis (Jia et al., 2009). Two main considerations must be taken into account when dealing with negation. First, the negation scope may affect only a word (***no reason***), a proposition (*Beckham* ***does not want to play again for Real***) or even a subject (***No one*** *would like to do*

[3] http://gate.ac.uk/

[4] http://nlp.stanford.edu/software/lex-parser.shtml

this). Different approximations have been proposed to delimit the scope of negation. Some assume the scope to be those words between the negation token and the first punctuation mark (Pang et al., 2002), others consider a fixed number of words after the negation token (Hu and Liu, 2004). Second, the impact of negation is usually neutralized by reversing the polarity of the sentence (Polanyi and Zaenen, 2006) or using contextual valence shifters which increase or dismiss the final value of negativity or positivity of the sentence (Kennedy and Inkpen, 2006).

In this work, the negation scope is detected using the syntax tree and dependencies generated by the Stanford Parser. The dependency *neg* allows us to easily determine the presence of several simple types of negation, such as those preceded by *don't*, *didn't*, *not*, *never*, etc. Other words not identified with this dependency, but also with a negation meaning, such as *no, none¸ nor* or *nobody*, are identified using a negation token list. To determine the negation scope, we find in the syntax tree the first common ancestor that encloses the negation token and the word immediately after it, and assume all descendant leaf nodes to be affected by the negation.

For each concept in the sentence that falls into the scope of a negation, the system retrieves its antonym emotional category, if any, and assigns this category to the concept. If no antonym emotion is obtained, the concept is labeled with no emotion, according to the premise that the negation may change or neutralize the emotional polarity. An example of this process can be shown in the sentence "*Children and adults enamored of all things pokemon won't be disappointed*". In this sentence, the Stanford Parser discovers a negation and the system, through the syntax tree, determines that the scope of the negation encloses the words "*won't be disappointed*". As the synset of "*disappointed*" has been labeled with the emotional category *despair*, its antonym is retrieved, and the emotional category of the antonym, *hope,* is used to label the concept.

On the other hand, the quantifiers are words considered in sentiment analysis as *amplifiers* or *downtoners* (Quirk et al., 1985). That is to say, the word *very* in the sentence "*That is a very good idea"* amplifies the intensity of the emotional meaning and the positivity of the sentence, while the word *less* in the sentence "*It is less handsome than I was expecting*" dismisses its intensity and polarity. The most common approach to identify quantifiers is the use of lists of words which play specific grammatical roles in the sentence. These lists normally contain a fixed value for all positive words and another value for all negative words (Polanyi and Zaenen, 2006). By contrast, Brooke (2009) proposes a novel approach where each quantifier is assigned its own polarity and weight.

The quantifiers are usually represented as sentence modifiers, assuming their scope to be the whole sentence and modifying its overall polarity. However, when dealing with sentences like "*The house is really nice and the neighborhood is not bad*", these approaches assume that the quantifier *really* amplifies the intensity of both conjunctives, when it only should amplify the intensity of the first one. By contrast, our approach determines the scope of the quantifiers by the syntax tree and the dependencies over them. Thus, when a quantifier is detected in a sentence, the dependencies are checked and only those that play certain roles, such as adverbial or adjectival modifiers, are considered. All concepts affected by a quantifier are marked with the weight corresponding to that quantifier, which will serve to amplify/dismiss the emotions of these concepts in the classification step. The quantifier list used here is the one proposed in Brooke (2009).

The sentence "*Stale first act, scrooge story, blatant product placement, some very good comedic songs*" illustrates the analysis of the quantifiers. The system detects two tokens which are in the quantifier list and play the appropriate grammatical roles. The first quantifier *some* affects to the words "*very good comedic songs*", while the second quantifier *very* only affects to "*good*". So these concepts are marked with the specific weight of each quantifier. Note that the concept "*good*" is marked twice.

4.4 Sentence classification

Up to this point, the sentence has been labeled with a set of emotional categories, negations and their scope have been detected and the quantifiers and the concepts affected by them have been identified. In this step, this information is used to translate the sentence into a *Vector of Emotional Occurrences (VEO)*, which will be the input to the machine learning classification algorithm. Thus, each sentence is represented as a vector of 16 values, each of one representing an emotional category. The VEO vector is generated as follows:

- If the concept has been labeled with an emotional category, the position of the vector for this category is increased in *1*.

- If no emotional category has been found for the concept, then the category of its first hypernym labeled is used. As the hypernym generalizes the meaning of the concept, the value assigned to the position of the emotional category in the VEO is weighted as follows:

$$VEO[i] = VEO[i] + \frac{1}{Hyper.Depth + 1}$$

- If a negation scope encloses the concept, then the antonym emotion is used, as described in the previous step. The emotional category position of this antonym in the VEO is increased in *0.9*. Different tests have been carried out to set this parameter, and the *0.9* value got the best results. The reason for using a lower value for the emotional categories derived from negations is that the authors consider that a negation changes the emotional meaning of a concept but usually in a lower percentage. For example, the sentence "*The neighborhood is not bad*" does not necessarily mean that it is a good neighborhood, but it is a quite acceptable one.
- If a concept is affected by a quantifier, then the weight of that quantifier is added to the position in the VEO of the emotional category assigned to the concept.

Thus, a sentence like "*This movie…. isn't worth the energy it takes to describe how really bad it is*" will be represented by the VEO [1.0, 0, 0.0, 0, 0, 0.0, 0, 0, 2.95, 0, 0, 0, 0, 0, 0, 0]. In this sentence, the concept *movie* is labeled with the emotional category *joy*, the concept *worth* is labeled with *positive-expectation*, the concept *energy* is labeled with *liking,* and the concept *bad* is labeled with *dislike.* Since the concepts *worth* and *energy* fall into the negation scope, they both change their emotional category to *dislike.* Besides, since the concept *bad* is amplified by the quantifier *really*, the weight of this concept in the VEO is increased in *0.15*.

5 Evaluation framework and results

In this work, two different corpora have been used for evaluation (see Section 3): a movie review corpus containing 1000 sentences labeled with either a positive or negative polarity; and a news headlines corpus composed of 1000 sentences labeled with an emotional intensity value between -100 and 100.

To determine the best machine learning algorithm for the task, 20 classifiers currently implemented in Weka[5] were compared. We only show the results of the best performance classifiers: a logistic regression model *(Logistic)*, a C4.5 decision tree (*J48Graph)* and a support vector machine (*LibSVM).* The best outcomes for the three algorithms were reported when using their default parameters, except for J48Graph, where the confidence factor was set to *0.2*. The evaluation is accomplishes using 10-fold cross validation. Therefore, 100 instances of each dataset are held back for testing in each fold, and the additional 900 instances are used for training.

5.1 Evaluating polarity classification

We first analyze the effect of expanding the coverage of the emotional lexicon by semi-automatically adding to WordNet Affect more synsets labeled with emotional categories, as explained in Section 3. To this end, we compare the results of the method using three different affective lexical databases: WordNet Affect and WordNet Affect extended with 372 and 603 synsets, respectively. For the sake of comparing the results in both corpora, the news dataset has been mapped to a -100/100 classification (-100 = [-100, 0), 100 = [0,100]).

Table 2 shows the results as average precision and accuracy of these experiments. Note that, as the weight of mislabeling for both classes is the same and the classes are balanced, accuracy is equal to recall in all cases. Labeling 975 new synsets significantly improves the performance of our system in both datasets and for all ML techniques. In particular, the best improvement is achieved by the Logistic classifier: from 52.7% to 72.4% of accuracy in the news dataset, and from 50.5% to 61.5% of accuracy in the movies dataset.

Emotional Lexicon	**Method**	**News Corpus**		**Movie Reviews**	
		Pr.	***Ac.***	***Pr.***	***Ac.***
WNAffect	*Logistic*	52.8	52.7	51.3	50.5
	J48Graph	27.7	52.6	50	50
	LibSVM	27.7	52.6	53.2	50.6
WNAffect + 372 synsets	*Logistic*	69.9	65.2	53.9	53.8
	J48Graph	70.1	64.8	55.3	55.1
	LibSVM	68.9	63.9	52	51.8
WNAffect + 603 synsets	*Logistic*	73.8	72.4	61.6	61.5
	J48Graph	73.6	70.9	60.9	60.9
	LibSVM	71.6	70.3	62.5	59.4

Table 2: Precision and accuracy percentages achieved by our system using different affective databases.

[5] http://www.cs.waikato.ac.nz/ml/weka/

We have observed that, especially in the news dataset, an important number of sentences that are labeled with a low positive or negative emotional intensity could be perfectly considered as neutral. The intensity of these sentences highly depends on the previous knowledge and particular interpretation of the reader. For instance, the sentence "*Looking beyond the iPhone*" does not express any emotion itself, unless you are fan or detractor of Apple. However, this sentence is labeled in the corpus with a *15* intensity value. It is likely that these kinds of sentences introduce noise into the dataset. In order to estimate the influence of such sentences in the experimental results, we conducted a test removing from the news dataset those instances with an intensity value in the range [-25, 25]. As expected, the accuracy of the method increases substantially, i.e. from 72.4% to 76.3% for logistic regression.

The second group of experiments is directed to evaluate if dealing with negations and quantifiers improves the performance of the method. To this end, the approach described in Section 4.3 was applied to both datasets. Table 3 shows that processing negations consistently improves the accuracy of all algorithms in both datasets; while the effect of the quantifiers is not straightforward. Even if we expected that using quantifiers would lead to better results, the performance in both datasets decreases in 2 out of the 3 ML algorithms. However, combining both features improves the results in both datasets. The reason seems to be that, when no previous negation detection is performed, if the emotional category assigned to certain concepts are wrong (because these concepts are affected by negations), the quantifiers will be weighting the wrong emotions.

Features	**Method**	**News Corpus**		**Movie Reviews**	
		Pr.	***Ac.***	***Pr.***	***Ac.***
Negations	*Logistic*	74.2	72.5	61.7	61.6
	J48Graph	74.1	71.2	62.8	62.6
	LibSVM	72.7	71.1	62.4	60.1
Quantifiers	*Logistic*	73.7	72.2	61.9	61.9
	J48Graph	73.6	70.9	59.5	59.5
	LibSVM	72.1	70.6	61.1	59
Negations + Quantifiers	*Logistic*	74.4	72.7	62.4	62.4
	J48Graph	74.1	71.2	62.5	62.1
	LibSVM	72.8	71.2	62.6	60.5

Table 3: Precision and accuracy of the system improved with negation and quantifier detection.

The comparison with related work is difficult due to the different datasets and methods used in the evaluations. For instance, Pang et al. (2002) use the Movie Review Polarity Dataset, achieving an accuracy of 82.9% training a SVM over a bag of words. However, their aim was to determine the polarity of documents (i.e. the whole movie reviews) instead of sentences. When working at the sentence level, the information from the context is missed, and the results are expected to be considerably lower. As a matter of fact, it happens that many sentences in the Sentence Polarity Movie Review Dataset are labeled as positive or negative, but do not express any polarity when taken out of the context of the overall movie review. This conclusion is also drawn by Meena and Prabhakar (2007), who achieve an accuracy of 39% over a movie review corpus (not specified) working at the sentence level, using a rule based method to analyze the effect of conjuncts. This accuracy is well below that of our method (62.6%).

Molianen and Pulman (2007) present a sentiment composition model where the polarity of a sentence is calculated as a complex function of the polarity of its parts. They evaluate their system over the SemEval 2007 news corpus, and achieve an accuracy of 65.6%, under our same experimental conditions, which is also significantly lower than the accuracy obtained by our method.

5.2 Evaluating intensity classification

Apart from identifying of polarity, we also want to examine the ability of our system to determine the emotional intensity in the sentences. To this aim, we define two intensity distributions: the 3-classes and the 5-classes distribution. For the first distribution, we map the news dataset to 3-classes: negative [-100, -50), neutral [-50, 50) and positive [50, 100]. For the second distribution, we map the dataset to 5-classes: strongly negative [-100, -60), negative [-60, -20), neutral [-20, 20), positive [20, 60) and strongly positive [60, 100]. We can see in Table 4 that, as the number of intensity classes increases, the results are progressively worse, since the task is progressively more difficult.

Intensity classes	**Method**	**News Corpus**	
		Pr.	***Ac.***
2-classes	*Logistic*	74.4	72.7
	J48Graph	74.1	71.2
	LibSVM	72.8	71.2
3-classes	*Logistic*	60.2	63.8
	J48Graph	66	64.8
	LibSVM	54.8	64.6
5-classes	*Logistic*	48.3	55.4
	J48Graph	47.3	54.8
	LibSVM	43.1	53.1

Table 4: Precision and accuracy in three different intensity classification tasks.

The 3-classes distribution coincides exactly with that used in one of the SemEval 2007 Affective task, so that we can easily compare our results with those of the systems that participated in the task. The CLaC and CLaC-NB systems (Andreevskaia and Bergler, 2007) achieved, respectively, the best precision and recall. CLaC reported a precision of 61.42 % and a recall of 9.20%; while CLaC-NB reported a precision of 31.18% and a recall of 66.38%. Our method clearly outperforms both systems in precision, while provides a recall (which is equal to the accuracy) near to that of the best system. Besides, our results for both metrics are well-balanced, which does not occur in the other systems.

Regarding the 5-classes distribution evaluation, to the authors' knowledge no other work has been evaluated under these conditions. However, our system reports promising results: using 5 classes it achieves better results than other participant in the SemEval task using just 3 classes (Chaumartin, 2007; Katz et al., 2007).

5.3 Evaluating the effect of word ambiguity on sentiment analysis

A further test has been conducted to examine the effect of word ambiguity on the classification results. To this aim, we repeated the experiments above without using WSD. First, we simply assigned to each word its first sense in WordNet. Second, we selected these senses randomly. The results are presented in Table 5. We only show those of the best algorithm for each intensity distribution.

Intensity classes	**Method**	**News Corpus**	
		Pr.	***Ac.***
2-classes (Logistic)	*WSD*	74.4	72.6
	1st Sense	71.6	69.3
	Random Sense	69.1	64.1
3-classes (J48Graph)	*WSD*	66	64.8
	1st Sense	59	62.9
	Random Sense	50.8	61
5-classes (Logistic)	*WSD*	48.3	55.4
	1st Sense	43.7	53.8
	Random Sense	46.8	51.6

Table 5: Precision and accuracy for three different word disambiguation strategies.

It can be observed that, even though the use of word disambiguation improves the classification precision and accuracy, the improvement with respect to the first sense heuristic is less than expected. This may be due to the fact that the senses of the words in WordNet are ranked according to their frequency, and so the first sense of a word is also the most frequent one. Besides, the Most Frequent Sense (MFS) heuristic in WSD is usually regarded as a difficult competitor. On the contrary, the improvement with respect to the random sense heuristic is quite remarkable.

6 Conclusions and future work

In this paper, a novel approach to sentence level sentiment analysis has been described. The system has resulted in a good method for sentence polarity classification, as well as for intensity identification. The results obtained outperform those achieved by other systems which aim to solve the same task.

Nonetheless, some considerations must be noted. Even with the extended affective lexicon, around 1 in 4 sentences of each corpus has not been assigned any emotional category, sometimes because their concepts are not labeled in the lexicon, but mostly because their concepts do not have any emotional meaning *per se*. A test on the news corpus removing those sentences not labeled with any emotional meaning has been performed for the 2-classes classification problem, allowing the method to obtain an accuracy of 81.7%. However, to correctly classify these sentences, it would be necessary to have additional information about their contexts (i.e. the body of the news item, its section in the newspaper, etc.).

Finally, the authors plan to extend the method to deal with modal and conditional operators, which will allow us to distinguish among situations that have happened, situations that are happening, situations that *could*, *might* or *possibly* happen or will happen, situations that are wished to happen, etc.

Acknowledgments

This research is funded by the Spanish Ministry of Science and Innovation (TIN2009-14659-C03-01), the Comunidad Autonoma de Madrid and the European Social Fund through the IV PRICIT program, and the Spanish Ministry of Education through the FPU program.

References

Julian Brooke. 2009. A Semantic Approach to Automated Text Sentiment Analysis. Simon Fraser University. Ph. D. Thesis.

Jorge Carrillo de Albornoz, Laura Plaza and Pablo Gervás. 2010. Improving Emotional Intensity Classification using Word Sense Disambiguation. *Research in Computing Science 46* :131-142.

François-Régis Chaumartin. 2007. UPAR7: A Knowledge-based System for Headline Sentiment Tagging. In *Proceedings of the 4th Workshop on Semantic Evaluations (SemEval 2007),* pages 422-425.

Ann Devitt and Khurshid Ahmad. 2007. Sentiment Polarity Identification in Financial News: A Cohesion-based Approach. In *Proceedings of the 45th Annual Meeting of the ACL,* pages 984-991.

Andrea Esuli and Fabrizio Sebastiani. 2006. Determining Term Subjectivity and Term Orientation for Opinion Mining. In *Proceedings of the 11th Conference of the EACL,* pages 193-200.

Minging Hu and Bing Liu. 2004. Mining and Summarizing Customer Reviews. In *Proceedings of the 10th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 168-177.

Lifeng Jia, Clement Yu and Weiji Meng. 2009. The Effect of Negation on Sentiment Analysis and Retrieval Effectiveness. In *Proceeding of the 18th ACM Conference on Information and Knowledge Management*, pages 1827-1830.

Phil Katz, Matthew Singleton and Richard Wicentowski. 2007. SWAT-MP: the SemEval-2007 Systems for Task 5 and Task 14. In *Proceedings of the 4th Workshop on Semantic Evaluations (SemEval 2007),* pages 308-313.

Alistair Kennedy and Diana Inkpen. 2006. Sentiment Classification of Movie Reviews Using Contextual Valence Shifters. *Computational Intelligence 22(2):* 110-125.

Soo-Min Kim and Eduard Hovy. 2004. Determining the Sentiment of Opinions. In *Proceedings of COLING 2004,* pages 1367-1373.

Justin Martineau and Tim Finin. 2009. Delta TFIDF: An Improved Feature Space for Sentiment Analysis. In *Proceedings of the 3rd AAAI International Conference on Weblogs and Social Media.*

Arun Meena and T.V. Prabhakar. 2007. Sentence Level Sentiment Analysis in the Presence of Conjuncts Using Linguistic Analysis. In *Proceedings of ECIR 2007,* pages 573-580.

George A. Miller, Richard Beckwith, Christiane Fellbaum Derek Gross and Katherine Miller. 1990. Introduction to WordNet: An On-Line Lexical Database. *International Journal of Lexicography 3(4)*:235-244.

Karo Moilanen and Stephen Pulman. 2007. Sentiment Composition. In *Proceedings of RANLP 2007,* pages 378-382.

Roser Morante and Walter Daelemans. 2009. A Metalearning Approach to Processing the Scope of Negation. In *Proceedings of the CONLL 2009*, pages 21-29.

Bo Pang, Lillian Lee and Shivakumar Vaithyanathan. 2002. Thumbs up? Sentiment Classification using Machine Learning Techniques. In *Proceedings of CoRR 2002.*

Bo Pang and Lillian Lee. 2004. A Sentimental Education: Sentiment Analysis using Subjectivity Summarization based on Minimum Cuts. In *Proceedings of the 42nd Annual Meeting of the ACL,* pages 271-278.

Siddharth Patwardhan, Satanjeev Banerjee and Ted Pedersen. 2005. SenseRelate::TargetWord - A Generalized Framework for Word Sense Disambiguation. In *Proceedings of the ACL 2005 on Interactive Poster and Demonstration Sessions*, pages 73-76.

Livia Polanyi and Annie Zaenen. 2006. Contextual Valence Shifters. Computing Attitude and Affect in Text: Theory and Applications. In *The Information Retrieval Series 20,* pages 1-10.

Randolph Quirk, Sidney Greenbaum, Geoffrey Leech and Jan Svartvik. 1985. A Comprehensive Grammar of the English Language. Longman.

Carlo Strapparava and Alessandro Valitutti. 2004. Wordnet-Affect: an Affective Extension of WordNet. In *Proceedings of the LREC 2004*, pages 1083-1086.

Peter D. Turney. 2002. Thumbs up or Thumbs down?: Semantic Orientation applied to Unsupervised Classification of Reviews. In *Proceedings of the 40th Annual Meeting of the ACL,* pages 417-424.

Casey Whitelaw, Navendu Garg and Shlomo Argamon. 2005. Using Appraisal Groups for Sentiment Analysis. In *Proceedings of the 14th ACM Conference on Information and Knowledge Management,* pages 625-631.

Janyce M. Wiebe, Rebecca F. Bruce and Thomas P. O'Hara. 1999. Development and Use of a Goldstandard Data Set for Subjectivity Classification. In *Proceedings of the 37th Annual Meeting of the ACL,* pages 246-253.

Theresa Wilson, Janyce Wiebe and Paul Hoffman. 2005. Recognizing Contextual Polarity in Phrase-level Sentiment Analysis. In *Proceedings of the HLT-EMNLP 2005*, pages 347-354.