Journal of Universal Computer Science, vol. 18, no. 16 (2012), 2319-2333
submitted: 9/6/11, accepted: 27/8/12, appeared: 28/8/12 © J.UCS

Learning to Classify Neutral Examples from Positive and Negative Opinions

María-Teresa Martín-Valdivia
(Department of Computer Science. University of Jaén, Spain
maite@ujaen.es)

Arturo Montejo-Ráez
(Department of Computer Science. University of Jaén, Spain
amontejo@ujaen.es)

Alfonso Ureña-López
(Department of Computer Science. University of Jaén, Spain
laurena@ujaen.es)

Mohammed Rushdi Saleh
(Department of Computer Science. University of Jaén, Spain
msaleh@ujaen.es)

Abstract: Sentiment analysis is a challenging research area due to the rapid increase of subjective texts populating the web. There are several studies which focus on classifying opinions into positive or negative. Corpora are usually labeled with a star-rating scale. However, most of the studies neglect to consider neutral examples. In this paper we study the effect of using neutral sample reviews found in an opinion corpus in order to improve a sentiment polarity classification system. We have performed different experiments using several machine learning algorithms in order to demonstrate the advantage of taking the neutral examples into account. In addition we propose a model to divide neutral samples into positive and negative ones, in order to incorporate this information into the construction of the final opinion polarity classification system. Moreover, we have generated a corpus from Amazon in order to prove the convenience of the system. The results obtained are very promising and encourage us to continue researching along this line and consider neutral examples as relevant information in opinion mining tasks.

Keywords: Opinion Mining, Sentiment Polarity, Neutral Examples, NLP
Categories: I.2.7, I.7, I.2.1, H.3.3, L.3.2

1 Introduction

Recently the interest in Sentiment Analysis (SA) and Opinion Mining (OM), has grown significantly due to various different factors [Liu, 2010]. The rapid evolution of the World Wide Web has changed our view of the Internet. It has turned into a collaborative framework where technological and social trends come together, resulting in the over exploited term Web 2.0. In addition, the tremendous use of e-commerce services has been accompanied by an increase in freely available online reviews and opinions about products and services. A customer who wants to buy a

product usually searches for information on the Internet trying to find other consumer analyses. In fact, web sites such as Amazon, Epinions or IMDb (Internet Movie Database), can greatly affect customer´s decisions. Moreover, opinion mining is useful not only for the individual customer but also for any company or institution as a powerful tool for understanding customer preferences. However, the huge amount of information available makes it necessary to develop new methods and strategies.

SA is becoming one of the main research areas in Natural Language Processing (NLP) and Text Mining (TM). This new discipline attempts to identify and analyze opinions and emotions [Tsytsarau and Palpanas, 2012]. It includes several subtasks such as subjectivity detection [Wiebe et al., 2001], polarity classification [Pang et al., 2002], review summarization [Somprasertsri and Lalitrojwong, 2010], humor detection [Mihalcea and Strapparava, 2006], emotion classification [Strapparava and Mihalcea, 2008] and so on. Specifically, this paper focuses on sentiment polarity classification.

Different approaches have been applied in the field of sentiment polarity classification, but there are two main trends: In the symbolic approach, which applies manually crafted rules and lexicons, the document is represented as a collection of words. Then the sentiment of each word can be determined by different methods, for example, using a web search [Hatzivassiloglou and Wiebe, 2000] or consulting lexical resources such as WordNet [Kamps et al., 2004]. The other approach relies on machine learning techniques to tackle the classification of reviews according to their orientation (positive or negative). In this approach, the document is represented by different features and a machine learning algorithm is applied. These features may include the use of n-grams or defined grammatical roles like adjectives, for instance. Commonly used machine learning algorithms include Support Vector Machines (SVM), Maximum Entropy (ME) or Naïve Bayes (NB) [Pang et al., 2002].

This paper focuses on a particular issue regarding the opinion polarity at document level: the use of neutral examples in order to classify the review as positive or negative. We train a classifier using a corpus labeled with a numerical rating for each opinion. In the first step we only use the positive and negative reviews to train the system. With this model we classify the neutral examples into positive or negative samples and then include them in the corpus in order to train the final classifier.

We use different machine learning algorithms in order to classify the polarity of reviews. Specifically we use Support Vector Machine, Logistic Regression and K Nearest Neighbors. We focus on how neutral opinions can be included in order to improve the classification of sentiment polarity. We tested different combinations of neutral examples with the positive and negative sets, and even without using any neutral review. Furthermore, we developed a method for classifying the neutral examples into positive or negative reviews. In our experiments we used different corpora labeled according to the rating of each review. The paper is organized as follows: Section 2 briefly describes previous related work on sentiment polarity classification and discusses how neutral samples can affect this challenging task. In Section 3 the data sets used in our experiments are described. We then explain the methodology used and describe the three machine learning algorithms applied in our experiments, along with the experimental framework developed. Section 5 presents the experiments carried out and discusses the main results obtained. Finally, we outline conclusions and further work.

2 Use of neutral examples in sentiment polarity

Nowadays, sentiment polarity is one of the main tasks in opinion mining. Given a subjective text, a sentiment polarity classifier must determine whether the opinion is positive or negative. In the scenario of commercial product reviews, it would be interpreted as if the customer likes (positive) or dislikes (negative) a given product overall. The opinions can be ranked into a specific ranking between 1 and 5 stars or between 1 and 10. Moreover, sentiment polarity classification can be studied at document, sentence or feature level. Document level polarity classification attempts to classify the general sentiments into reviews, news, or articles [Wiebe et al., 2001; Pang et al., 2002; Mullen and Collier, 2004], while sentence-level polarity classification tries to determine the sentiment for each sentence [Yi et al., 2003; Pang and Lee, 2005], and feature level tries to find different sentiments within one sentence [Wilson et al., 2005]. Some systems classify the opinions detected using different scales [Pang and Lee, 2008]. In some cases, the sole purpose is to identify opinions in a text and classify them into positive, negative or neutral classes. In other cases, the goal is to assign different ratings such as very bad, bad, satisfactory, good, very good, or excellent.

There are a variety of rating systems in the web and blogs which include opinions and reviews of products and services. The simplest one solely includes a binary classification of the reviews (positive or negative, thumbs up or thumbs down). Other sites use a star-based rating or numerical system (1 to 5 stars for example in Amazon, or 1 to 10 points in the IMDb).

There are different ways to treat the neutral examples in the corpus. For example, in a 5-star rating system, some studies neglected the neutral examples in the corpus. Thus, the reviews rated with 1 and 2 stars were classified as negative while 4 and 5 were labeled as positive [Turney, 2002; Pang et al., 2002; Dave et al., 2003; Yu and Hatzivassiloglou, 2003]. In this case, reviews labeled with 3 stars (i.e., neutral examples) are not included in the learning process. The information supplied by the 3 star opinions is simply disregarded. However, there are some papers showing how the use of neutral examples can help to improve the classification [Pang and Lee, 2005]. For example, [Koppel and Schler, 2006] suggest that the polarity problem might be best handled as a three-class problem with positive, negative and neutral classes. Moreover, they conclude that the use of neutral training examples in learning facilitates better distinction between positive and negative opinions.

In addition to rating systems, some web sites include other useful information about the reviewed item such as recommended and non-recommended products (for example, Epinions). Usually, 1 and 2 star reviews are labeled as non-recommended and 4 and 5 stars are labeled as recommended. However, for the 3 star reviews we can find opinions that sometimes are labeled as recommended and other reviews as non-recommended. In this type of corpus, this additional information classifying opinions as positive or negative can avoid the noise introduced by the 3 star reviews. Unfortunately, this kind of corpus is not common and usually it is necessary to decide what to do with the 3 star samples. This is a very difficult problem even for human users who must decide the polarity of neutral examples because some of them tend to be positive while others have a negative orientation. In Figure 1 and Figure 2 we can see two 3 star reviews from the Amazon site. We have underlined the positive

sentences, and the negative text is in bold. The first review tends to be positive and the second one seems to be negative. However, this is only the user's subjective appraisal. Therefore, in this work we will study the effect of using neutral examples to train a classifier using a machine learning approach. Our proposal is to incorporate the information supplied by neutral samples in order to train a classifier and improve a sentiment polarity system.

I bought this camera while i was pregnant because i Fig.d i would need a good one for when the baby came. I was really pleased with it and it did take really nice photos. **The videos werent the best** but i suppose you cant expect perfect videos from a cheap camera. When the baby came i had someone running around the delivery room snapping pictures. **Every other picture was blurry.** I dont know if it was the operator or just the camera. **I did notice that you had to wait a long time and have the perfect light for the camera to take really good pics.** After having the camera for about 2 or 3 months i had an accident involving dirty baby clothes a misplaced camera and a washing machine...needles to say the camera didnt make it out alive. I decided to go ahead and buy the same camera again. I was still pleased with it but a little bummed i couldnt find the 10mp for as cheap so instead i had to settle for the 8.2. **Anyways im an avid review reader and i had read a couple that said the camera straight up quit working after 6 months.** I decided to ignore them because most of the other comments were totally positive. I had my second camera for about 5 months and it died...on its own...no washing machine involved. **It was like the Auto Focus just completely quit working for some reason.** (really bad timing too because i was at the hospital with my friend while she was having HER baby when i found out it quit working.) Anyway i liked the camera but cant decide if i want to try it out a

Figure 1: Example of a 3 star review with positive orientation

This is a nice camera if you're looking primarily for a camera that is small, rugged, and waterproof**. However, if you're looking for a camera that takes great pictures - keep looking. The image quality is terrible so forget the 10.1 mega pixel feature. And since the 3.6 optical zoom is hardly enough to zoom in on far away objects the poor image quality becomes a big deal. When you crop a photo in an attempt to "zoom" digitally, you can see terrible pixilation, grain, and blur. I considered sending the camera back to Amazon, but decided to keep it for taking photos in the water. If I didn't want that feature I would have definitely returned it for something else.**

Figure 2: Example of a 3 star review with negative orientation

3 Corpora description

In this paper we have used different corpora. Firstly, we performed several experiments with the Taboada corpus in order to demonstrate that the correct use of neutral examples can improve the sentiment polarity classification system. Then we trained a classifier using the 3 star samples in our SINAI corpus, demonstrating the advantages of taking the neutral examples into account. We briefly describe the two corpora in the next subsections.

#Stars	#Reviews
1	80
2	88
3	20
4	51
5	132
Total	**371**

Table 1: Review in the Taboada corpus according to the number of stars

3.1 Taboada corpus

This collection was used by [Taboada and Grieve, 2004] and by [Taboada et al., 2006] with the main goal of classifying text automatically based on subjective content. They applied a standard method for calculating semantic orientation by extracting the adjectives. This method is based on [Turney, 2002] where the combinations of adjective + noun and noun + noun were used. The corpus includes 400 opinions collected from the website Epinions.com divided into 200 reviews classified as "recommended" (positive) and 200 as "non-recommended" (negative). The texts contain opinions about products and services like movies, books, cars, cookware, phones, hotels, music and computers. The total number of categories is eight and the corpus contains 25 positive and 25 negative reviews for each category.

Although the reviews in the Epinions website use a 5-star rating system, the available Taboada corpus only includes opinions labeled with "recommended" and "non-recommended" tags, and the reviews are not rated with the number of stars. For this reason we asked the Taboada research group to supply us with the original corpus that they had crawled from the Internet in order to work with a star rating system. Hence we received 371 files because some files were missing from the source. Table 1 shows the distribution of reviews in the Taboada corpus according to the number of stars.

In this corpus all the 1 and 2 star reviews are also labeled as "non-recommended", while 4 and 5 star opinions are tagged as "recommended". As regards the 20 reviews with 3 stars, 14 of them are tagged as "non-recommended" and 6 "recommended". So the whole collection includes 182 (168+14) negative samples and 189 (183+6) positive reviews.

3.2 SINAI corpus

Unfortunately most of the opinion corpora published do not include the labels "recommended" and "non-recommended", so it is necessary to decide what to do with the neutral examples. Many authors simply neglect the 3 star reviews and only work with clearly positive and negative samples in the corpora. However, some studies show that the correct use of neutral examples significantly improves the polarity classification systems, as commented previously [Koppel and Schler, 2006]. Thus it is very interesting to study the best way to include the 3 star examples in our systems. For this reason we generated our own corpus, called SINAI, by crawling the Amazon

website, and it is freely available for the scientific community through http://sinai.ujaen.es/wiki/index.php/SINAISaCorpus. SINAI stands for the name of our research group "Sistemas INteligentes de Acceso a la Información" (Intelligent Systems for Information Access). In order to build the corpus we extracted opinions about cameras of different brands and series. A total of 1,942 documents were labeled with different numbers of stars. Table 2 shows the distribution of reviews per camera model.

Camera Model	**#Reviews**
CanonA590IS	400
CanonA630	300
CanonSD1100IS	426
KodakCx7430	64
KodakV1003	95
KodakZ740	155
Nikon5700	119
Olympus1030SW	167
PentaxK10D	126
PentaxK200D	90
Total	**1,942**

Table 2: Number of reviews per product in the SINAI corpus

#Stars	**#Reviews**
1	78
2	67
3	96
4	411
5	1,290
Total	**1,942**

Table 3: Reviews in the SINAI corpus according to the number of stars

The opinions in Amazon are rated using a 5-star scale, but they do not include additional information about recommended or non-recommended items. Table 3 shows the distribution of reviews according to the number of stars for the SINAI corpus. In the same way as in the Taboada corpus we used 1 and 2 stars as negative samples and 4 and 5 stars as positive reviews. However, the 3-star reviews must be treated in a different way.

The original SINAI corpus is also extremely unbalanced, with the number of positive reviews (rated with 4 and 5 stars) clearly higher than the number of negative reviews (rated with 1 and 2 stars). So we randomly chose 200 positive examples from the total of positive reviews. The new corpus also contains the 145 negative reviews and the 96 neutral examples. This corpus has been called SINAI-B (SINAI Balanced corpus) and was built with the sole purpose of testing the effect of neutral examples on a balanced corpus that does not include the "recommended" and "not recommended" information for each review.

4 Methodology

In this section we describe the framework followed in our experiments, mainly based on the training of different classifiers in order to determine the polarity of reviews in an opinion corpus. Specifically, we applied the Support Vector Machine (SVM), Logistic Regression (LR) and K-Nearest Neighbor (KNN).

4.1 Machine Learning Algorithms

The SVM algorithm [Vapnik, 1995] has been applied successfully in many text classification tasks due to these features [Joachims, 1998]: first, it is robust in high dimensional spaces; second, any feature is relevant; third, it works well when there is a sparse set of samples; finally, most text categorization problems are linearly separable. In addition, SVM has achieved good results in opinion mining, and this algorithm has surpassed other machine learning techniques [O'Keefe and Koprinska, 2009].

Logistic Regression (LR) is a mathematical modeling approach in which the best-fitting, yet least-restrictive model is desired to describe the relationship between several independent explanatory variables and a dependent dichotomous response variable. Some studies have been successful applying this model in the area of sentiment analysis [Martínez-Cámara et al., 2011].

K-Nearest Neighbors (KNN) is a case-based learning method, which keeps all the training data for classification. KNN is very simple; for each new item to be classified KNN seeks the k closest items in the training set, and then it returns the major class in the "neighbors" set. KNN has been used in other opinion mining studies, obtaining good results [Tan and Zhang, 2008].

4.2 Experimental framework

We have used the Rapid Miner software with its text mining plug-in which contains different tools designed to assist in the preparation of text documents for mining tasks (tokenization, stop word removal and stemming, among others). Rapid Miner is an environment for machine learning and data mining processes that is freely available from htpp://rapid-i.com.

As regards the document model, we used the Vector Space Model (VSM) in order to generate the bag of words for each document. The English Porter stemming algorithm was applied in order to reduce words to their common root or stem. We also

removed some tokens using a stop word list. However, we did preserve some useful sentiment information such as “ok” and “not”.

For SVM, we implemented our experiments using the libsvm learner by [Chang and Lin, 2001], which is integrated into Rapid Miner as one of the available functions. In our experiments we applied a Linear SVM with the default configuration set by the tool (C-SVC type, RBF kernel, epsilon equal to 0.001 and shrinking heuristics enabled). For LR we used the kernel type Anova available in Rapid Miner with the default values for the other parameters. Finally, for KNN we used the Euclidean distance (1-NN) because it is the configuration with the best results.

4.3 Experiments

Our experiments were run on the Taboada corpus and SINAI corpus. They are different in domain and size. The Taboada corpus contains eight categories with different domains, while the SINAI corpus includes nine different models of cameras (thus, only one domain). In order to train the classifier the corpus is divided into positive and negative samples. For both corpora we considered reviews with 1 and 2 stars as negative samples and reviews with 4 and 5 stars as positive ones. However, for the 3 star reviews we performed different partitions, and thus several training corpora were generated:

- N12P45: the 3 star reviews were ignored
- N123P45: the 3 star examples are considered as negative reviews
- N12P345: the 3 star examples are considered as positive reviews

In addition, the Taboada corpus includes information about recommended and non-recommended items, so we can use this important information to train the classifier. Thus, we included the 3 stars labeled “non-recommended” in the negative set and the 3 stars tagged with “recommended” in the positive samples (N12NR3P45R3). Unfortunately, reviews expressed in most of the opinion forums do not include the recommended and non-recommended information, and only the number of stars for each review is supplied. This is the case with the SINAI corpus. In this situation, it is necessary to develop a method to decide about the polarity of 3 star reviews. Thus we generated a model using the training data from the SINAI corpus but excluding the 3 star reviews and only using 1, 2, 4 and 5 star opinions. We obtained a preliminary classifier C1 which we used to classify the 3 star examples. We used this new classification and we added the new positive 3 star reviews to the 4 and 5 star set and the new negative 3 star opinions to the 1 and 2 star negative set. This new corpus including the neutral examples was used to generate a completely new classifier C2. Figure 3 shows the process followed to generate our classifier. The experiments performed following this strategy have been called N12CNR3P45CRP3.

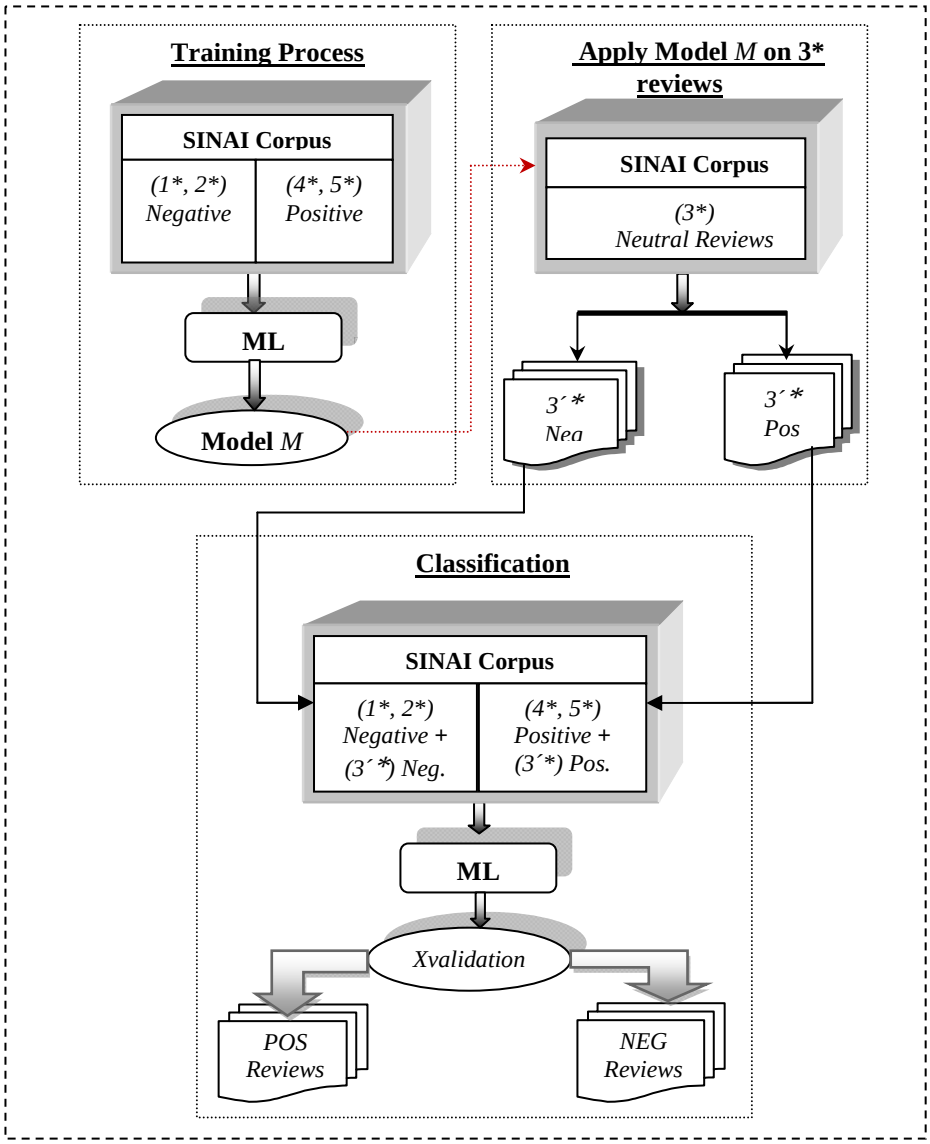

Figure 3: Process followed to build the final classifier C2

4.4 Evaluation

The system has been evaluated by applying 10-fold cross validation on each corpus, and measuring performance according to the indicators given below:

$$Precision\ (P) = \frac{tp}{tp + fp} \quad (1)$$

$$Recall\ (R) = \frac{tp}{tp + fn} \quad (2)$$

$$Accuracy = \frac{tp + tn}{tp + tn + fn + fp} \quad (3)$$

$$k = \frac{\Pr(a) + \Pr(e)}{1 - \Pr(e)} \quad (5)$$

$$Pr(a) = \frac{tp + tn}{tp + tn + fn + fp} \quad (6)$$

$$Pr(e) = \frac{(tp + fp)(fn + tn)(tp + fn)(fp + tn)}{(tp + tn + fn + fp)^2} \quad (7)$$

where *tp* (True Positives) are those assessments where the system and human expert agree on a label assignment, *fp* (False Positives) are those labels assigned by the system which do not agree with the expert assignment, *fn* (False Negatives) are those labels that the system failed to assign as they were given by the human expert, and *tn* (True Negatives) are those non assigned labels that were also discarded by the expert [see Tab. 4]. The precision tells us how well the labels are assigned by our system (the fraction of assigned labels that are correct). The recall measures the fraction of expert labels found by the system. Finally, accuracy combines both precision and recall, calculating the proportion of true results (both true positives and true negatives). *k*: Kappa; *Pr(a)* is the relative observed agreement among raters; *Pr(e)* is the hypothetical probability of chance agreement [Sebastiani, 2002].

	True Yes	True No
Predicted Yes	*tp*	*fp*
Predicted No	*fn*	*tn*

Table 4: Contingency table

5 Results and discussion

The experiments were divided into three parts: the first one was run on the Taboada corpus, the second one on the original SINAI corpus and the third part on the SINAI-B corpus. For each corpus we performed experiments for each partition using different combinations of neutral examples. As a reminder, the corpora that have been used are:

- N12P45: the 3 star reviews were ignored
- N123P45: the 3 star examples are considered as negative reviews
- N12P345: the 3 star examples are considered as positive reviews
- N12NR3P45R3 (only applicable to Taboada corpus): the 3 stars labeled "non-recommended" are included in the negative set and the 3 stars tagged with "recommended" are considered as positive samples.
- N12CNR3P45CRP3: the corpus includes the 3 star examples tagged as "recommended" by the C1 classifier into the positive samples and the 3 star examples tagged as "non-recommended" by the C1 classifier into the negative samples.

In addition, these experiments were run using the three machine learning algorithms SVM, LR and KNN.

The experiments accomplished with the Taboada corpus are shown in Table 5. As presumed, the best result was obtained when recommended and non-recommended information in the 3 star reviews (N12NR3P45R3) was taken into account. The 20 reviews labeled with 3 stars were distributed between 6 as positive (recommended) and 14 as negative (non-recommended). However, it is very interesting to note that the second best results were obtained when we applied the approach described in Figure 3 for all the algorithms (N12CNR3P45CRP3). According to the machine learning algorithm, LR clearly overcomes the other two algorithms. In addition, the Kappa measure is also bigger for LR than for SVM and KNN.

Regarding the SINAI corpus, we performed almost the same experiments as with the Taboada corpus, except for the case where the "recommended" and "non-recommended" information was used. Table 6 shows the results obtained. Although the best results were achieved with the new model proposed in Figure 3, the improvement is not as significant as the one obtained with Taboada corpus. We think the main reason for this is the high accuracy already obtained with the baseline case. This makes it very difficult to improve the final results. In fact, the best improvement is obtained with KNN, the algorithm with the worst accuracy. Nevertheless, the experiments reinforce our hypothesis about the improvement when neutral examples are used.

Algorithm	Corpus	Precision	Recall	Accuracy	Kappa
SVM	N12P45	78.62%	87.37%	80.38%	0.603
	N123P45	79.05%	85.88%	81.66%	0.634
	N12P345	76.25%	85.71%	77.10%	0.531
	N12NR3P45R3	81.84%	91.49%	85.16%	0.702
	N12CNR3P45CRP3	80.74%	88.05%	**82.74%**	0.653
LR	N12P45	87.39%	86.20%	86.32%	0.725
	N123P45	86.22%	85.69%	85.70%	0.714
	N12P345	85.25%	84.03%	84.39%	0.639
	N12NR3P45R3	87.52%	87.31%	87.33%	0.746
	N12CNR3P45CRP3	88.24%	87.57%	**87.61%**	0.751
KNN	N12P45	72.34%	71.27%	71.53%	0.427
	N123P45	79.05%	75.88%	71.66%	0.634
	N12P345	71.67%	70.00%	70.88%	0.405
	N12NR3P45R3	74.91%	69.57%	72.90%	0.393
	N12CNR3P45CRP3	74.21%	72.19%	**72.80%**	0.449

Table 5: Taboada corpus with different distribution of 3 star reviews

Algorithm	Corpus	Precision	Recall	Accuracy	Kappa
SVM	N12P45	94.38%	99.65%	94.20%	0.421
	N123P45	92.02%	98.77%	91.41%	0.489
	N12P345	94.70%	99.28%	94.19%	0.413
	N12CNR3P45CRP3	94.64%	99.61%	**94.44%**	0.466
LR	N12P45	93.68%	68.26%	94.80%	0.497
	N123P45	88.89%	68.47%	91.50%	0.481
	N12P345	91.28%	63.26%	94.24%	0.374
	N12CNR3P45CRP3	95.45%	69.87%	**95.01%**	0.536
KNN	N12P45	63.66%	64.11%	80.38%	0.273
	N123P45	63.80%	65.20%	84.19%	0.286
	N12P345	63.41%	64.49%	80.49%	0.276
	N12CNR3P45CRP3	68.10%	70.23%	**89.03%**	0.377

Table 6: SINAI corpus with different distribution of 3 star reviews

As imbalance may affect classifier behavior, a drawback of the original SINAI corpus is the great difference between the number of positive and negative examples. So we performed the same experiments with the SINAI-B corpus. The results are shown in Table 7 and as we can see the results obtained when we consider the neutral examples are better than when we neglect them, although in this case the improvement is slightly lower than with the original SINAI corpus. However, these experiments highlight the advantage of using the neutral examples in an appropriate way.

6 Conclusions

This paper focuses on the importance of neutral examples in reviews used in sentiment polarity classification tasks. We have applied several machine learning algorithms on different corpora in order to classify the sentiment polarity of subjective documents. We proposed a model to divide the neutral examples of a corpus into positive and negative samples. This information was then incorporated into the original corpus in order to regenerate and improve the model.

Algorithm	Corpus	Precision	Recall	Accuracy	Kappa
SVM	N12P45	84.69%	90.50%	84.92%	0.687
	N123P45	79.62%	77.47%	80.72%	0.610
	N12P345	80.18%	89.89%	78.23%	0.474
	N12CNR3P45CRP3	87.54%	83.40%	**86.17%**	0.722
LR	N12P45	93.30%	92.67%	93.04%	0.857
	N123P45	87.81%	87.52%	87.73%	0.752
	N12P345	83.60%	78.47%	83.47%	0.604
	N12CNR3P45CRP3	93.33%	93.17%	**93.20%**	0.862
KNN	N12P45	74.18%	72.88%	73.94%	0.460
	N123P45	70.75%	70.40%	70.52%	0.406
	N12P345	63.66%	63.48%	67.35%	0.267
	N12CNR3P45CRP3	75.19%	75.31%	**75.28%**	0.502

Table 7: SINAI-B corpus with different distribution of 3 star reviews

The results obtained encourage us to continue working along this line. Thus, in future work we will include more information on neutral examples in order to improve the classification, for example, using external resources like SentiWordNet [Baccianella et al., 2010]. In addition, we will apply the classifier developed to other corpora such as the Pang corpus on movie reviews [Pang and Lee, 2008].

Acknowledgments

This work has been partially funded by the European Commission under the Seventh (FP7-2007-2013) Framework Programme for Research and Technological Development through the FIRST project (FP7-287607). This publication reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein. It has been also partially supported by a grant from the Fondo Europeo de Desarrollo Regional (FEDER) through the TEXT-COOL 2.0 project (TIN2009-13391-C04-02) from the Spanish Government. Also another part of this project was funded by Agencia Española de Cooperación Internacional para el Desarrollo MAEC-AECID.

References

[Baccianella et al., 2010] Baccianella, S., Esuli, A., & Sebastiani, F. (2010). Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In Proceedings

of LREC-10, 7th Conference on Language Resources and Evaluation, Valletta, MT, 2010, pages 2200-2204.

[Chang and Lin, 2001] Chang, C. C., Lin, & C. J. (2001). LIBSVM: a library for support vector machines. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm. (last accessed 01/02/2011)

[Dave et al., 2003] Dave, K., Lawrence, S., & Pennock, D. M. (2003). Mining the peanut gallery: opinion extraction and semantic classification of product reviews. In WWW '03: Proceedings of the 12th international conference on World Wide Web. ACM, New York, NY, USA, pp. 519–528.

[Hatzivassiloglou and Wiebe, 2000] Hatzivassiloglou, V., & Wiebe, J. (2000). Effects of adjective orientation and gradability on sentence subjectivity. In: COLING-00: Proceedings International Conference on Computational Linguistics. pp. 299–305.

[Joachims, 1998] Joachims, T. (1998). Text categorization with support vector machines: learning with many relevant features. In: N´edellec, C., Rouveirol, C. (Eds.), Proceedings of ECML-98, 10th European Conference on Machine Learning. No. 1398. Springer Verlag, Heidelberg, DE, Chemnitz, DE, pp. 137–142.

[Kamps et al., 2004] Kamps, J., Marx, M., Mokken, R. J., & Rijke, M. D. (2004). Using wordnet to measure semantic orientation of adjectives. In Proceedings of LREC-04, Conference on Language Resources and Evaluation, pp. 1115–1118.

[Koppel and Schler, 2006] Koppel, M., & Schler, J. (2006). The importance of neutral examples for learning sentiment. Computational Intelligence 22 (2), 100–109.

[Liu, 2010] Liu, B. (2010). Sentiment Analysis and Subjectivity, Handbook of Natural Language Processing, Second Edition.

[Martínez-Cámara et al., 2011] Martínez-Cámara, E., Martín-Valdivia, M.T.,. Ureña-López, L.A. (2011) Opinion Classification Techniques Applied to a Spanish Corpus. Proceedings of Natural Language Processing and Information Systems, pp. 169-176.

[Mihalcea and Strapparava, 2006] Mihalcea, R.,Strapparava, C.: Learnin to Laugh (automatically): Computational Models for Humor Recognition, Journal of Computational Intelligence, Vol. 22, 2006, 126-142

[Mullen and Collier, 2004] Mullen, T., & Collier, N. (2004). Sentiment analysis using support vector machines with diverse information sources. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 412–418.

[O'Keefe and Koprinska, 2009] O'Keefe, T., & Koprinska, I. (2009). Feature selection and weighting methods in sentiment analysis. In: Proceedings of the 14th Australasian Document Computing Symposium, Sydney, Australia.

[Pang and Lee, 2004] Pang, B., & Lee, L. (2004). A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In: Proceedings of the ACL'04 Association for Computational Linguistics. pp. 271–278.

[Pang and Lee, 2005] Pang, B., Lee, L. (2005). Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In: Proceedings of the ACL'05 Association for Computational Linguistics. pp. 115–124.

[Pang and Lee, 2008] Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. Found. Trends Information Retrieval 2 (1-2), 1–135.

[Pang et al., 2002] Pang, B., Lee, L., Vaithyanathan, S. (2002). Thumbs up? Sentiment classification using machine learning techniques. In: Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 79–86.

[Sebastiani, 2002]. Sebastiani, F. (2002). Machine Learning in Automated Text Categorization. ACM Computing Surveys, 34(1), 1.

[Somprasertsri and Lalitrojwong, 2010] Somprasertsri, G., Lalitrojwong, P.: Mining Feature-Opinion in Online Custumer Reviews for Opinion Summarization, Journal of Universal Computer Science, vol. 16, 2010, 938-955.

[Strapparava and Mihalcea, 2008] Strapparava, C., Mihalcea, R.: Learning to identify emotions in text. In *Proceedings of the 2008 ACM symposium on Applied computing* (SAC '08). ACM, 2008, 1556-1560

[Taboada et al., 2006] Taboada, M., Anthony, C., & Voll, K. (2006). Methods for creating semantic orientation dictionaries. In Proceedings of 5th International Conference on Language Resources and Evaluation (LREC).

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

[Tan and Zhang, 2008] Tan, S., Zhang, J. (2008) .An empirical study of sentiment analysis for Chinese documents. Expert System with Applications 34, 2622–2629

[Tsytsarau and Palpanas, 2012] Tsytsarau, M. and Palpanas, T. (2012) Survey on mining subjective data on the web Data Mining and Knowledge Discovery. Vol. 24 (3) pp. 478-514. Doi: 10.1007/s10618-011-0238-6

[Turney, 2002] Turney, P. D. (2002). Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In: ACL '02: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics, Morristown, NJ, USA, pp. 417–424.

[Vapnik, 1995] Vapnik, V. N. (1995). The Nature of Statistical Learning Theory. Springer, New York.

[Wiebe et al., 2001] Wiebe, J., Wilson, T., Bell, M. (2001). Identifying collocations for recognizing opinions. In: Proceedings of the ACL'01 Association for Computational Linguistics Workshop on Collocation: Computational Extraction, Analysis, and Exploitation. pp. 24–31.

[Wilson et al., 2005] Wilson, T., Wiebe, J., & Hoffmann, P. (2005). Recognizing contextual polarity in phrase-level sentiment analysis. In: HLT '05: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Morristown, NJ, USA, pp. 347–354.

[Yi et al., 2003] Yi, J., Nasukawa, T., Bunescu, R., & Niblack, W. (2003). Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques. In: IEEE Intl. Conf. on Data Mining (ICDM). pp. 427–434.

[Yu and Hatzivassiloglou, 2003] Yu, H., & Hatzivassiloglou, V. (2003). Towards answering opinion questions: separating facts from opinions and identifying the polarity of opinion sentences. In: Proceedings of the 2003 conference on Empirical methods in natural language processing. Association for Computational Linguistics, Morristown, NJ, USA, pp. 129–136.