A Comparative Study of Polarity Lexicons to Identify Extreme Opinions

Sattam Almatarneh[1] and Pablo Gamallo[1]

Centro Singular de Investigación en Tecnoloxías da Información (CITIUS),
Universidad de Santiago de Compostela
Rua de Jenaro de la Fuente Domínguez, Santiago de Compostela 15782, Spain
sattam.almatarneh@usc.es,
pablo.gamallo@usc.es

1

Abstract. This paper describes a method to automatically build a sentiment lexicon, as well as a study comparing it with four well-known sentiment lexicons. For this purpose, an indirect evaluation is carried out. The lexicons are integrated into supervised sentiment classifiers and their performance is evaluated in two sentiment classification tasks in order to identify i) the most negative vs. not most negative opinions, and ii) the most positive vs. not most positive. Moreover, a set of textual features is integrated into the classifiers so as to analyze how these textual features improve the lexicon performance.

Keywords: Sentiment Analysis, Opinion Mining, Sentiment Lexicon, Linguistic Features, Polarity Classification, Extreme Opinion.

2 Introduction

In the last decade, a huge number of studies have been carried out in the field of opinion mining which is also called sentiment analysis. The motivation behind these studies was the need to extract useful information to be used in many domains from the vast amount of available users' views on blogs, social networks, news, and shopping websites. At the forefront of all fields, business intelligence is the most attractive domain for opinion mining. Most work in this domain focus on mining customers' reviews for better market understanding. Another traditional field is government intelligence, where the focus is on many issues such as elections, parties' reputation, and choosing policies according to people opinions.

The fundamental task in Opinion Mining is polarity classification [15], which occurs when a piece of text stating an opinion is classified into a predefined set of polarity categories (e.g., positive, neutral, negative). Reviews such as "thumbs up" versus "thumbs down", or "like" versus "dislike" are examples of two-class polarity classification.

An unusual way of performing sentiment analysis is to detect and classify extreme opinions, which represent the most negative and most positive opinions about a topic, an object or an individual. An extreme opinion is the worst or the best view, judgment, or appraisal formed in one's mind about a particular matter.

Extreme opinions only constitute a small portion of the opinions on Social Media. According to [19], only about 5% of all opinions are on the most extreme points of a scale. However, extreme views have a mighty impact on product sales considering they guide customer decisions before buying. The experiments reported in[14] analyzed this relationship, which found that as the high proportion of negative online consumer reviews increased, the consumer's negative attitudes also increased. Similar effects have been observed in consumer reviews: one-star reviews significantly hurt book sales on Amazon.com [6]. The impact of 1-star reviews, which represent the most negative views, is greater than the impact of 5-star reviews in this particular market sector.

One of the main motivations for detecting extreme opinions is the fact that they actually stand for *pure* positive and negative opinions. As rating systems have no clear borderlines on a continuum scale, weakly polarized opinions (e.g. those rated as 4 and 2 in a 1 to 5 rating system) may be in fact closer to neutral statements. According to Pang and Lee [19], "it is quite difficult to properly calibrate different authors' scales, since the same number of *stars* even within what is ostensibly the same rating system can mean different things for different authors". Given that rating systems are defined on a subjective scale, only extreme opinions can be seen as natural, transparent, and non ambiguous positive / negative statements.

The main objective of this article is to examine the effectiveness of the automatic construction of a sentiment lexicon using an indirect evaluation procedure. The indirect evaluation consists of measuring the performance of supervised machine learning classifiers based on the lexicon. Our main contribution is to report an extensive set of experiments aimed to compare our automatic construction lexicon with other four well-known handcraft lexicons for two binary classification tasks:

- very negative (MN) *vs.* not very negative opinions (NMN)
- very positive (MP) *vs.* not very positive opinions (NMP)

Furthermore, a set of textual features is integrated into the classifiers to analyze how these textual features improve the lexicons performance in each task.

The rest of the paper is organized as follows. In the following section (3), we introduce some related work. Then, Section 4 describes the method. Experiments are introduced in Section 5, where we also describe the evaluation and discuss the results. We draw the conclusions and future work in Section 6.

3 Related Work

There are two main approaches to find sentiment polarity at a document level. First, supervised techniques, second, unsupervised strategies based on polarity lexicons. In recent years, many surveys and books in sentiment analysis describing the main methods and comparing the usefulness of different linguistic and textual features have appeared, such as [10,15].

Sentiment words also called opinion words are considered the primary building block in sentiment analysis as it is an essential resource for most sentiment analysis algorithms, and the first indicator to express positive or negative opinions. There are, at least, two ways of building sentiment lexicons: hand-craft elaboration [24,18,12,7], and automatic construction on the basis of an external resource. The automatic strategy builds the sentiment lexicons using diverse resources. Two different automatic strategies may be identified according to the nature of these resources: thesaurus and corpora.

[8] described the creation of two corpus-based lexicons. First, a general lexicon using SentiwordNet and the Subjectivity Lexicon. Second, a domain-specific lexicon using a corpus of drug reviews depending on statistical information. [17] built a lexicon containing a combination of sentiment polarity (positive, negative) with one of eight possible emotion classes (anger, anticipation, disgust, fear, joy, sadness, surprise, trust) for each word.

We propose a new method to create sentiment lexicons from multiple domains for extreme opinions, namely the most negative and most positive words. As far as we know, this is quite a different resource with regard to existing lexicons.

In spite of a vast number of existing approaches, a limited number of studies have offered an explicit comparison between sentiment analysis methods. [9] presented comparisons of eight popular sentiment analysis methods in terms of coverage and agreement. They developed a new method that combines existing approaches, providing the best coverage results and competitive agreement. [23] introduced a comparison of twenty-four popular sentiment analysis methods at the sentence-level, based on a benchmark of eighteen labeled datasets. The performance has been evaluated in two sentiment classification tasks: two classes (negative *vs.* positive) and three classes (negative, neutral and positive). However, these studies did not compare the efficiency of sentiment analysis methods or sentiment lexicons in the specific task of identifying extreme opinions. To the best of our knowledge, except for our study reported in [2], there is no previous work focused on the identifiation of this kind of opinions.

4 Method

We deal with two document-level binary classification tasks: 1) very negative *vs.* not very negative, and 2) very positive *vs.* not very positive. These tasks can be performed by using classifiers modeled with training data in a supervised strategy. Some linguistic characteristics of documents will be encoded as features in vector representation. These vectors and the corresponding labels feed the

classifiers. In the experiments described later, we will examine the following two types of features: sentiment lexicons and textual properties.

4.1 Sentiment Lexicons

In our previous studies [3,1], we described a strategy to build sentiment lexicons from corpora with standard positive and negative words. In this study, by contrast, we will use the same method to create two extreme lexicons: one containing the most negative words and the other consisting of the most positive ones. The method of constructing this corpus-based lexicon is summarized as follows: The first step to create lexicons is to measure the relative frequency (RF) for every word w in each category c according to equation 1:

$$RF_c(w) = \frac{freq(w,c)}{Total_c} \tag{1}$$

where c is any category of the star rating, from 1 to N; $freq(w,c)$ is the number of tokens of the target word in c; and $Total_c$ is the total number of word tokens in c. As in our experiments the corpus was PoS tagged, words are actually represented as (word, tag) pairs. Besides, we only work with adjectives and adverbs as they are the most relevant part of speech in sentiment analysis for any language, according to [4].

The second step is to calculate the average of RF values for two ranges of categories: most negative (MN) *vs* not most negative (NMN), and most positive (MP) *vs* not most positive (NMP). For this purpose, it is necessary to define a borderline value B for extreme opinions, which might vary according to the specific star rating of the reviews. For instance, if the rating goes from 1 to 10, and the borderline value B=2, the MN reviews are considered those rated from 1 to 2, while MP are those rated from 8 to 10. This is similar if the rating goes from 1 to 5 and the borderline is set at 1. In this case, the MN reviews are considered those rated 1, while MP are those rated 5. Given a borderline value, B, the average of the MN scores, $AvMN$, for a word is computed as follows:

$$AvMN(w) = \frac{\sum_{c=1}^{B} RF_c(w)}{B} \tag{2}$$

On the other hand, given $R = N - B$, where N is the total number of categories, the average of NMN values, $AvNMN$, for each word is computed in equation 3:

$$AvNMN(w) = \frac{\sum_{c=B+1}^{N} RF_c(w)}{R} \tag{3}$$

As for the average of MP scores, $AvMP$, for a word, it is computed in equation 4:

$$AvMP(w) = \frac{\sum_{c=(N+1)-B}^{N} RF_c(w)}{B} \tag{4}$$

And the average of NMP values, $AvNMP$, for each word is computed in equation 5:

$$AvNMP(w) = \frac{\sum_{c=1}^{N-B} RF_c(w)}{R} \tag{5}$$

In the following step, the objective is to assign polarity weights to words and classify them by using four polarity classes: MN, NMN, MP, and NMP. Extreme words (MN and MP) are separated from not extreme words by just comparing the difference between the average values obtained by the equations defined above: 2, 3, 4, 5. With this simple idea, we build two lexicons: one lexicon on the negative scale from MN to NMN, and another lexicon on the positive scale from MP to NMP. So, given a word w, we compute the differences D_{neg} and D_{pos} in equations 6 and 7, and assign the resulting values to w:

$$D_{neg}(w) = AvNMN(w) - AvMN(w) \tag{6}$$
$$D_{pos}(w) = AvNMP(w) - AvMP(w) \tag{7}$$

D_{neg} gives a weight to w within the negative scale, while D_{pos} assigns weights in the positive ranking. These two weights are used to classify words in the four aforementioned categories and thereby building two new polarity lexicons, which we call *VERY-NEG* and *VERY-POS*. Classification is carried out with the following basic algorithm:

If the value of $D_{neg}(w)$ is negative, w is in the MN class. If $D_{neg}(w)$ is positive, w is in NMN. If the value of $D_{pos}(w)$ is positive, w is in the MP class. If $D_{pos}(w)$ is negative, w is in NMP.

4.2 Set of Textual Features (SOTF)

Many textual features may be used as evidences to detect extreme views: both very positive or very negative alike. In this study, we have extracted some of them to examine to what extent they influence the identification of extreme views. Uppercase characters may indicate that the writer is very upset or affected, so we counted the number of words written in uppercase letters. Also, intensifier words could be a reliable indicator of the existence of extreme views. So, we considered words such as *mostly, hardly, almost, fairly, really, completely, definitely, absolutely, highly, awfully, extremely, amazingly, fully*, and so on.

Furthermore, we took into account negation words such as no, not, none, nobody, nothing, neither, nowhere, never, etc. In addition, we also considered elongated words and repeated punctuation such as(*sooooo, baaaaad, woooow, gooood, ???, !!!!,...etc.*). These textual features have been shown to be effective in many studies related to polarity classification such as [24,13].

Table 1 summarizes all the features introduced above with a brief description for each one.

Features	Descriptions
Lexicons (4 feat.)	Number and proportion of MN terms in the documents Number and proportion of NMN terms in the documents
SOTF (8 Feat.)	Number and proportion of negation words in the document Number and proportion of uppercase words in the document Number and proportion of elongated words and punctuations in the document Number and proportion of intensifiers words in the document

Table 1. Description of all the linguistic features.

5 Experiments

In order to cover several domains, the experiments were carried out using different datasets, including books, DVD, electronics, and housewares reviews. In our experiments, we automatically built two polarity lexicons using the strategy defined above in Subsection 4.1. Our lexicons were evaluated and compared with other existing handcraft lexicons in the two tasks of classifying reviews. Before defining the evaluation protocol and showing the results, we describe the resources, both lexicons and corpus-based datasets, used in the experiments.

5.1 Lexicons

As mentioned earlier, there are many popular and available sentiment lexicons. There are two types: First, lexicons assigning PoS tags to lemmas, such as SO-CAL and SentiWords. In our experiments, only adjectives and adverbs were compared. Second, lexicons without POS tags: Opinion Lexicon and AFINN-111.

Six lexicons will be compared depending on each task:: the two lexicons we automatically built using our strategy, called VERY-NEG and VERY-POS, and four manual resources: SO-CAL [24], SentiWords [7], Opinion Lexicon [12,16], and AFINN-111 [18].

VERY-NEG and VERY-POS Our proposed lexicons were built from the text corpora introduced in [22]. The corpora[1] consist of online reviews collected from IMDB, Goodreads, OpenTable and Amazon/Tripadvisor. Each of the reviews in this collection has an associated star rating: one star (most negative) to ten stars (most positive) in IMDB, and one star (most negative) to five stars (most positive) in all the other corpora.

Reviews were tagged using the Stanford Log-Linear Part-Of-Speech Tagger. Then, tags were broken down into WordNet PoS Tags: a (adjective), n (noun), v (verb), r (adverb). Words whose tags were not part of those categories were filtered out. The list of selected words was then stemmed.

[1] http://www.stanford.edu/~cgpotts/data/wordnetscales/

Word	Tag	Category	Freq	Total
bad	a	1	1127	699695
bad	a	2	2595	2507147
bad	a	3	2859	4207700
bad	a	4	2544	7789649
bad	a	5	1905	8266564

Table 2. A sample of the collection format for the pair ("bad", *a*) in each category.

Table 2 shows quantitative information for the adjective "bad", where *Freq* is the total number of tokens of a (word,tag) pair in each category and corpus, while *Total* is the total number of word tokens in each category and corpus (Total values are constant for all words but repeated for each one in order to make processing easier). Then, we compute AvMN, AvNMN, AvMP and AvNMP for each word and obtain the weights (D_{neg} and D_{pos}) values to build the corresponding lexicons for each corpus. Finally, we compute the average of all weights for each word in order to obtain two cross-domain final lexicons[2] (VERY-NEG and VERY-POS). VERY-NEG contains a list of the most negative words (MN) and a list of words that are not classified as most negative (NMN). In the same way, VERY-POS contains two lists: the most positive words (MP) and the other words (NMP).

Through preliminary experiments, we found that the best results were obtained by filtering out words with very low weight ($D <= 0.00000001$), which are values close to zero. This means that we filtered out neutral words, i.e. words without polarity.

In order to ensure that all cases are tested, we created lexicons at two different borderline (B) values: B=1 and B=2. The former is used to determine extreme values on scales from 1 to 5. More precisely, when B=1 we mean that 1 (most negative) and 5 (most positive) are the extreme scores. The latter parametrization (B=2) is used to define extreme values in scales from 1 to 10: in this case, 1 and 2 are extreme values for most negatives, while 9 and 10 represent the class of most positive opinions. Each of our two lexicons, VERY-NEG and VERY-POS, consists of two lists derived from different values of B, as shown in Table 3.

[2] https://github.com/almatarneh/LEXICONS

Lexicon	Negative	Positive	Total	ADJ	ADV
VERY-NEG B=1	5270	8096	14460	11670	2790
VERY-NEG B=2	6232	2771	14328	11557	2771
VERY-POS B=1	5884	8287	14171	11402	2769
VERY-POS B=2	7092	7152	14244	11472	2772
SO-CAL	2005	1697	3702	2826	876
SentiWords	8152	8084	16236	13425	2811
Opinion Lexicon	4783	2007	6790	-	-
AFINN-111	1598	878	2476	-	-

Table 3. Number of words for each class negative and positive in each lexicon and total number of words (adjectives(ADJ) and adverbs(ADV)).

As our objective is to compare VERY-NEG and VERY-POS with other popular handcrafted lexical resources, we describe four existing lexicons in the next subsections. See Table 3.

SO-CAL Lexicon SO-CAL was described in [24]. The authors created their dictionary manually since they believe that the overall accuracy of lexicon-based sentiment analysis mainly relies on the quality of those resources. The lexicon was built with content words, namely adjectives, adverbs, nouns and verbs, adding sentiment scores between -5 and +5. The Negative sign (-) refers to negative polarity while the positive sign (+) indicates positive polarity, and any semantically neutral word has zero score.

SentiWords Lexicon Sentiwords[3] is a sentiment lexicon derived from SentiWordNet using the method described in [7]. It contains more than 16,000 words provided with a sentiment score between -1 (very negative) and +1 (very positive). The words in this lexicon are arranged with WordNet synsets, that include adjectives, nouns, verbs and adverbs.

Opinion Lexicon A list of negative and positive sentiment words for English[4]. 6790 words, 2007 positive words, and 4783 negative words. This list was accumulated across several years starting from the papers [12,16]. Includes mis-spellings, morphological variants, slang, and social-media mark-up.

AFINN-111 [18] has presented another manually generated lexicon called AFINN[5]. In this lexicon, a list of English words has been constructed and rated for valence with an integer between minus five (negative) and plus five (positive).

[3] http://hlt-nlp.fbk.eu/technologies/sentiwords
[4] https://www.cs.uic.edu/ liub/FBS/sentiment-analysis.html
[5] http://www2.imm.dtu.dk/pubdb/views/publication_details.php?id=6010

5.2 Multi-Domain Sentiment Dataset

This dataset[6] was used in [5]. It contains product reviews taken from Amazon.com for 4 types of products (domains): Kitchen, Books, DVDs, and Electronics. The star ratings of the reviews are from 1 to 5 stars. In our experiments, we adopted the scale with five categories. In this case, the borderline separating the MN values from the rest was set to 1, which stands for the MN reviews. The documents in the other four categories were put in the NMN class. According to this borderline value, the MP class was made up of those reviews scored with 5, while the NMP class was built with the rest of reviews. Table 4 shows the number of reviews in each class for each task.

Datasets	**# of Reviews**	**Negative**	**Positive**	**MN**	**NMN**	**MP**	**NMP**
Books	2000	1000	1000	532	1462	731	1269
DVDs	2000	1000	1000	530	1470	714	1286
Electronics	2000	1000	1000	666	1334	680	1320
Kitchens	2000	1000	1000	687	1313	754	1246

Table 4. Size of the four test datasets and the total number of reviews in each class negative *vs.* positive, (MN *vs.* NMN) and (MP *vs.* NMP)

5.3 Training and Test

Since we are facing a text classification problem, any existing supervised learning method can be applied. Support Vector Machines (SVM) has been shown to be highly effective at traditional text categorization [20]. We decided to utilize *scikit*[7] which is an open source machine learning library for Python programming language [21]. We chose SVM as our classifier for all experiments, hence, in this study we will only summarize and discuss results for this learning model. More specifically, we utilized the sklearn.svm.LinearSVC module[8]. Supervised classification requires two samples of documents: training and testing. The training sample will be used to learn various characteristics of the documents and the testing sample was used to predict and next verify the efficiency of our classifier in the prediction. The data set was randomly partitioned into training (75 %) and test (25 %). In all collections, the two-class categorization is unbalanced: much fewer MN and MP reviews than NMN and NMP ones. Therefore, as recommended in [11], we examined the performance by giving more importance to the positive class. We found that performance was sensitive to the SVM weights which modify the relative cost of misclassifying positive and negative samples. In our analysis, we employed 5_fold cross_validation and the effort was put on

[6] https://www.cs.jhu.edu/ mdredze/datasets/sentiment/index2.html
[7] http://scikit-learn.org/stable/
[8] http://scikit-learn.org/stable/modules/generated/sklearn.svm.LinearSVC.html

optimizing F1 which is computed with respect to MN and MP in the first two tasks (which is the target class):

$$F1 = 2 * \frac{P * R}{P + R} \tag{8}$$

where P and R are defined as follows:

$$P = \frac{TP}{TP + FP} \tag{9}$$

$$R = \frac{TP}{TP + FN} \tag{10}$$

Where TP stands for true positive, FP is false positive, and FN is false negative. To optimize F1, we tried out a grid search approach with exponentially growing sequences of the value of the parameter *class_weight*. More precisely, we tested class_weight with different values: $2^{-5}, 2^{-4}, 2^{-3}, 2^{-2}, ..., 2^{10}$. After finding the best value of class_weight within that sequence, we conducted a finer grid search on that better district (e.g. if the optimal value of class_weight is 8, then we test all the neighbors in this region: e.g. 3, 4, 5, 6, 7, 9, 10, 11, 12, 13, 14, 15 and 16). The class_weight was finally set to the value returning the highest F1 across all these experiments

5.4 Results

Tables 5 and 6 summarize the polarity classification results (in terms of (P, R, and F1) of the two classification tasks for all lexicons.

	BOOK			**DVD**			**Electronic**			**Kitchen**			
Lexicon	P	R	$F1$	P	R	$F1$	P	R	$F1$	P	R	$F1$	**Avg** ($F1$)
VERY-NEG B=1	0.46	0.76	0.58	0.57	0.64	**0.60**	0.52	0.86	**0.65**	0.53	0.72	0.61	**0.62**
VERY-NEG B=2	0.48	0.80	**0.60**	0.46	0.78	0.58	0.49	0.87	0.63	0.54	0.72	**0.62**	0.61
SO-CAL	0.44	0.64	0.52	0.45	0.73	0.56	0.55	0.71	0.62	0.43	0.92	0.58	0.58
SentiWorrds	0.41	0.66	0.51	0.42	0.66	0.52	0.54	0.67	0.60	0.45	0.93	0.61	0.57
Opinion Lexicon	0.42	0.66	0.52	0.48	0.80	**0.60**	0.50	0.85	0.63	0.44	0.94	0.60	0.60
AFINN-111	0.44	0.66	0.52	0.49	0.78	**0.60**	0.48	0.87	0.62	0.43	0.94	0.59	0.59
VERY-NEG B=1 +SOTF	0.48	0.75	0.59	0.57	0.62	0.60	0.53	0.86	**0.66**	0.53	0.72	0.61	**0.62**
VERY-NEG B=2 +SOTF	0.50	0.81	**0.62**	0.46	0.76	0.58	0.52	0.86	0.64	0.55	0.75	**0.64**	**0.62**
SO-CAL +SOTF	0.47	0.68	0.55	0.47	0.75	0.58	0.49	0.85	0.62	0.44	0.93	0.60	0.59
SentiWorrds +SOTF	0.44	0.66	0.53	0.44	0.68	0.53	0.48	0.84	0.61	0.45	0.93	0.61	0.57
Opinion Lexicon +SOTF	0.47	0.74	0.58	0.59	0.64	**0.61**	0.52	0.85	0.64	0.44	0.93	0.60	0.61
AFINN-111 +SOTF	0.47	0.69	0.56	0.61	0.61	**0.61**	0.51	0.90	0.65	0.45	0.92	0.61	0.61

Table 5. Polarity classification results for all collections with all lexicons, in terms of Precision (P), Recall (R), F1 scores and the average of all F1 for *most negative* class (MN). The best F1 in each dataset is highlighted (in bold).

Lexicon	BOOK			DVD			Electronic			Kitchen			Avg ($F1$)
	P	R	$F1$	P	R	$F1$	P	R	$F1$	P	R	$F1$	
VERY-POS B=1	0.56	0.80	**0.66**	0.47	0.85	0.60	0.51	0.65	0.57	0.52	0.73	0.60	**0.61**
VERY-POS B=2	0.57	0.78	**0.66**	0.45	0.78	0.57	0.50	0.69	0.58	0.46	0.89	**0.61**	**0.61**
SO-CAL	0.41	0.94	0.57	0.43	0.91	0.58	0.49	0.69	0.57	0.44	0.93	0.59	0.58
SentiWorrds	0.40	0.94	0.56	0.42	0.94	0.58	0.44	0.87	0.58	0.42	0.95	0.58	0.58
Opinion Lexicon	0.41	0.92	0.57	0.51	0.76	**0.61**	0.45	0.87	**0.60**	0.44	0.95	**0.61**	0.60
AFINN-111	0.40	0.93	0.56	0.43	0.91	0.58	0.42	0.88	0.57	0.44	0.92	0.60	0.58
VERY-POS B=1 +SOTF	0.58	0.80	**0.67**	0.47	0.83	0.60	0.52	0.70	**0.60**	0.52	0.80	**0.63**	**0.63**
VERY-POS B=2 +SOTF	0.58	0.77	0.66	0.45	0.78	0.57	0.52	0.71	**0.60**	0.52	0.77	0.62	0.61
SO-CAL +SOTF	0.44	0.93	0.59	0.44	0.89	0.59	0.44	0.86	0.58	0.47	0.89	0.61	0.59
SentiWorrds +SOTF	0.43	0.90	0.58	0.42	0.90	0.58	0.45	0.87	0.59	0.45	0.85	0.59	0.59
Opinion Lexicon +SOTF	0.44	0.88	0.59	0.52	0.75	**0.62**	0.46	0.85	0.59	0.46	0.91	0.61	0.60
AFINN-111 +SOTF	0.42	0.89	0.57	0.49	0.76	0.59	0.43	0.88	0.58	0.48	0.84	0.61	0.59

Table 6. Polarity classification results for all collections with all lexicons, in terms of Precision (P), Recall (R), F1 scores and the average of all F1 for *most positive* class (MP). The best F1 in each dataset is highlighted (in bold).

Considering the average of the four datasets (last column in Tables 5 and 6), the classifier configured with our lexicons outperforms the same classifier trained with the manual resources. The same thing happens when we add SOTF features to the classifier. However, it is worth noting that in two of the datasets, namely DVD and Electronic, the results seem more mitigated, which is going to require a deeper analysis of errors.

6 Conclusions

In this article, we have measure the quality of a corpus-based sentiment lexicon and some handcrafted resources by evaluating their performance in a supervised strategy to classify extreme opinions. The results of this indirect evaluation show that the automatically built lexicon has a stable behavior in different datasets and even improves other manually constructed resources. In future work, we will compare the performance of the same lexicons in an unsupervised method to classify extreme opinions.

References

1. Almatarneh, S., Gamallo, P.: Automatic construction of domain-specific sentiment lexicons for polarity classification. In: International Conference on Practical Applications of Agents and Multi-Agent Systems. pp. 175–182. Springer (2017)
2. Almatarneh, S., Gamallo, P.: Searching for the most negative opinions. In: International Conference on Knowledge Engineering and the Semantic Web. pp. 14–22. Springer (2017)
3. Almatarneh, S., Gamallo, P.: A lexicon based method to search for extreme opinions. PloS one 13(5), e0197816 (2018)

4. Benamara, F., Cesarano, C., Picariello, A., Recupero, D.R., Subrahmanian, V.S.: Sentiment analysis: Adjectives and adverbs are better than adjectives alone. In: ICWSM. Citeseer (2007)
5. Blitzer, J., Dredze, M., Pereira, F., et al.: Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In: ACL. vol. 7, pp. 440–447 (2007)
6. Chevalier, J.A., Mayzlin, D.: The effect of word of mouth on sales: Online book reviews. Journal of marketing research 43(3), 345–354 (2006)
7. Gatti, L., Guerini, M., Turchi, M.: Sentiwords: Deriving a high precision and high coverage lexicon for sentiment analysis. IEEE Transactions on Affective Computing 7(4), 409–421 (2016)
8. Goeuriot, L., Na, J.C., Min Kyaing, W.Y., Khoo, C., Chang, Y.K., Theng, Y.L., Kim, J.J.: Sentiment lexicons for health-related opinion mining. In: Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium. pp. 219–226. ACM (2012)
9. Gonçalves, P., Araújo, M., Benevenuto, F., Cha, M.: Comparing and combining sentiment analysis methods. In: Proceedings of the first ACM conference on Online social networks. pp. 27–38. ACM (2013)
10. Hemmatian, F., Sohrabi, M.K.: A survey on classification techniques for opinion mining and sentiment analysis. Artificial Intelligence Review pp. 1–51 (2017)
11. Hsu, C.W., Chang, C.C., Lin, C.J., et al.: A practical guide to support vector classification (2003)
12. Hu, M., Liu, B.: Mining and summarizing customer reviews. In: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 168–177. ACM (2004)
13. Kennedy, A., Inkpen, D.: Sentiment classification of movie reviews using contextual valence shifters. Computational intelligence 22(2), 110–125 (2006)
14. Lee, J., Park, D.H., Han, I.: The effect of negative online consumer reviews on product attitude: An information processing view. Electronic commerce research and applications 7(3), 341–352 (2008)
15. Liu, B.: Sentiment analysis and opinion mining. Synthesis lectures on human language technologies 5(1), 1–167 (2012)
16. Liu, B., Hu, M., Cheng, J.: Opinion observer: analyzing and comparing opinions on the web. In: Proceedings of the 14th international conference on World Wide Web. pp. 342–351. ACM (2005)
17. Mohammad, S.M., Turney, P.D.: Crowdsourcing a word–emotion association lexicon. Computational Intelligence 29(3), 436–465 (2013)
18. Nielsen, F.Å.: A new anew: Evaluation of a word list for sentiment analysis in microblogs. arXiv preprint arXiv:1103.2903 (2011)
19. Pang, B., Lee, L.: Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In: Proceedings of the 43rd annual meeting on association for computational linguistics. pp. 115–124. Association for Computational Linguistics (2005)
20. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up?: sentiment classification using machine learning techniques. In: Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10. pp. 79–86. Association for Computational Linguistics (2002)
21. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al.: Scikit-learn: Machine learning in python. Journal of machine learning research 12(Oct), 2825–2830 (2011)

22. Potts, C.: Developing adjective scales from user-supplied textual metadata. In: NSF Workshop on Restructuring Adjectives in WordNet. Arlington, VA (2011)
23. Ribeiro, F.N., Araújo, M., Gonçalves, P., Gonçalves, M.A., Benevenuto, F.: Sentibench-a benchmark comparison of state-of-the-practice sentiment analysis methods. EPJ Data Science 5(1), 23 (2016)
24. Taboada, M., Brooke, J., Tofiloski, M., Voll, K., Stede, M.: Lexicon-based methods for sentiment analysis. Computational linguistics 37(2), 267–307 (2011)