Automatic Construction of Domain-Specific Sentiment Lexicons for Polarity Classification

Sattam Almatarneh and Pablo Gamallo

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

Abstract. The article describes a strategy to build sentiment lexicons (positive and negative words) from corpora. Special attention will paid to the construction of a domain-specific lexicon from a corpus of movie reviews. Polarity words of the lexicon are assigned weights standing for different degrees of positiveness and negativeness. This lexicon is integrated into a sentiment analysis system in order to evaluate its performance in the task of sentiment classification. The experiments performed shows that the lexicon we generated automatically outperforms other manual lexicons when they are used as features of a supervised sentiment classifier.

Keywords: Sentiment Analysis, Opinion Mining, Sentiment Lexicon, Polarity Classification

1 Introduction

There exist two main approaches to finding the sentiment polarity at a document or sentence level. First, machine learning techniques based on training corpora annotated with polarity information and, second, strategies based on polarity lexicons. Lexicon-based approaches are very popular in sentiment analysis and opinion mining, and they play a key role in all applications in this field. The main concern of lexicon-based approaches is that most polarity words are domain-dependent since the subjectivity status of most words is very ambiguous. The same word may be provided with a subjective burden in a specific domain while it can refer to an objective information in another domain. It follows that domain-dependent lexicons should outperform general-purpose dictionaries in the task of sentiment analysis. However, the construction of domain-dependent polarity lexicons is a strenuous and boring task if it is made manually for each target domain. With the increasing of many sentiment corpora in diverse domains, the automatic generation of this kind of resources for many domains is becoming a fundamental task in opinion mining and sentiment analysis [4]. The objective of this article is to propose a method for automatically building polarity lexicons from corpora. More precisely, we focus on the construction of a domain-specific

lexicon from a corpus of movie reviews and its use in the task of sentiment analysis. The experiments reported in this article shows that our automatic resource outperforms other manual general-purpose lexicons when they are used as features of a supervised sentiment classifier.

The rest of the paper is organized as follows. In Section 2 we describe the related work. Then, Section 3 describes the method to create our proposed lexicon and how to use it in the classification task. The Experiments are introduced in section 4, where we also describe the evaluation and discuss the results. We draw conclusions in Section 5.

2 Related Work

There are, at least, two ways of building sentiment lexicons: hand-craft elaboration [13,5], and automatic construction on the basis of an external resource [9]. We are interested in the automatic strategy, which builds the sentiment lexicons using diverse resources. Two different automatic strategies may be identified according to the nature of these resources: thesaurus or corpora.

2.1 Thesaurus-based

This strategy requires seed sentiment words to bootstrap new polarity entries. They are based on the synonyms and antonyms structure of thesaurus. [6] report a thesaurus-based method that makes use of synonymy relation between adjectives in WordNet to generate a graph. More precisely, the authors measure the shortest path between the adjective and two basic sentiment seeds, "good" and "bad", to determine the polarity of a word. This is a semi-supervised learning method which starts with a lexical resource, WordNet, and a small list of seeds in order to expand the lexical resource in an iterative process. In a similar way, [7] propose a method that starts with three seed lists containing positive, negative and neutral words, which are also expanded with their synonyms in WordNet. Unlike these strategies, our method does not require any thesaurus to expand the lexicon with synonyms or antonyms.

2.2 Corpus-based

The work described in [14] is one of the pioneer studies focused on learning polarities from corpus by classifying reviews into two categories "recommend or not recommend" depending on the average number of positive and negative phrases appear in the review. Their algorithm consists of the following steps: first, it searches for phrases in the review by using a Part-Of-Speech (POS) tagger and then determines the polarity of the extracted phrases by computing Pointwise Mutual Information and Information Retrieval (PMI-IR). Then, the algorithm identifies those associative words returned by the search engine using the NEAR operator. Finally, the polarity of each phrase is determined by computing all the polarities returned by the search engine.

[8] present an automated approach for constructing a context-dependent lexicon from an unlabeled opinionated text collection based on existing lexicons and tagged consumer reviews. Each entry of this lexicon is a pair containing a sentiment term and different "aspect" terms associated with the former. The same sentiment term may diverge in polarity when co-occurring with a particular aspect term. This strategy is semi-supervised since it needs to start with a seed list of words or with an existing lexicon. By contrast, our method generates the lexicon of positive and negative adjectives and adverbs directly from any labeled corpus for any language without needs to start with the small set of words as a seed or any existing lexicon.

3 The Method

Our strategy consists of two tasks: first, we create a corpus-based polarity lexicon with two labels, negative and positive, and a polarity weight assigned to each word. Second, sentiment classification is performed by making use of this lexical resource.

3.1 Sentiment Lexicon Generation

We detail how to construct a lexicon that ranks words from the negative values to positive ones. The lexicon can be generated using any corpus of reviews labeled with star rating: one star (most negative) to N stars (most positive). The category set is the number of stars that can be assigned to the reviews. For instance, we are provided with 10 categories only if each review can be rated from 1 to 10. The first step to create our proposed lexicon 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 tags in sentiment analysis for any language, according to [2].

The second step is to calculate the average of the RF values for two ranges of categories: negative and positive. For this purpose, it is necessary to define two values: first, a borderline value for negative and positive opinions, which might vary according to the specific star rating of the reviews. Second, the number of neutral categories. For example, if the star rating goes from 1 to 10 categories and we set the borderline in 4 with two neutral categories, the negative reviews would be those rated from 1 to 4, while the positive reviews would be from 7 to 10. So the neutral reviews would be those rated from 5 to 6. Given a borderline

value, B, the average of the negative scores, Avn, for a word is computed as follows:

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

On the other hand, given Nt and N where N is the total number of categories, and Nt is the number of neutral categories, the average of positive scores, Avp, for each word is computed in equation 3:

$$Avp(w) = \frac{\sum_{c=B+Nt}^{N} RF_c(w)}{B} \tag{3}$$

In the following step, the negative and positive words are selected by comparing the values of Avn with Avp. Given a word w, we compute the difference D in equation 4 and assign this value to w, which stands for the final *weight* of the word:

$$D(w) = Avp(w) - Avn(w) \tag{4}$$

If the value of $D(w)$ is negative, w will be in the class of negative words. If the value of $D(w)$ is positive, w will be in the positive class.

3.2 Sentiment classification

As our aim is to evaluate the efficiency of our proposed lexicon, we train a sentiment classifier by making use of simple lexicon-based features, namely: the number of positive and negative terms in the document, and the proportion of positive and negative terms. We use just lexicon-based features because the purpose of the evaluation is to measure the quality of the given lexicon. Those features were used to train a Linear Support Vector Classifier (sklearn.svm.LinearSVC)[1] with the scikit-learn free software machine learning library for the Python programming language. Each dataset was randomly split into a training set and a test set (75% and 25% of the documents, respectively). The classifiers were optimised by applying 5-fold cross-validation against the training data.

4 Experiments

In our experiments, we automatically built a polarity lexicon using the strategy defined above in Section 3.1. Our lexicon was evaluated and compared with other two existing handcraft lexicons in the task of classifying reviews as positive or negative. For the purpose of evaluation, we used movie reviews.

Movie reviews have been examined for sentiment analysis and opinion mining in many studies [1,11]. We have chosen to deal with movie reviews in all experiments since many datasets are freely available in this domain. In addition to that, [14] found movie reviews is one of the most sensitive domains for sentiment

[1] http://scikit-learn.org/stable/modules/generated/sklearn.svm.LinearSVC.htmll

classification. The reason is that the negative opinions about a bad movie may contain positive words for describing the events or characters in the same movie. The contrary is also true. So, movie reviews are very challenging for sentiment analysis compared to other domains.

4.1 Lexicons

In the following, three lexicons will be compared: the lexicon we built using our strategy, called SPLM, a manual resource reported in [13], called SO-CAL, and SentiWords [3].

4.1.1 SPLM

Our proposed lexicon was built from the corpus introduced in [12]. The corpus[2] consists of data gathered from the user-supplied reviews at the IMDB. Each of the reviews in this collection has an associated star rating: one star (most negative) to ten stars (most positive). The reviews were tagged using the Stanford Log-Linear Part-Of-Speech Tagger. Then, tags were broken down into the WordNet Tags: a (adjective), n (noun), v (verb), r (adverb). Words whose tags were not part of those syntactic categories were filtered out. The list of selected words was then stemmed.

Word	Tag	Category	Count	Total
bad	a	1	122232	25395214
bad	a	2	40491	11755132
bad	a	3	37787	13995838
bad	a	4	33070	14963866
bad	a	5	39205	20390515
bad	a	6	43101	27420036
bad	a	7	46696	40192077
bad	a	8	42228	48723444
bad	a	9	29588	40277743
bad	a	10	51778	73948447

Table 1. A sample of the IMDB collection format for the word "bad" as adjective ("a") in each Category (from 1 to 10)

Table 1 shows a sample for the adjective "bad", where *Freq* is the total number of tokens of a (Word,Tag) pair in each Category (from rate 1 to 10), while *Total* is the total number of word tokens in each Category. Notice that Total values are constant for all words but they repeated for each one in order to make processing easier.

[2] http://compprag.christopherpotts.net/code-data/imdb-words.csv.zip

The next step is to compute *Avn* and *Avp* for each word. By making use of the equations defined above (3, 2 and 4), we obtain the weights assigned to each word-tag pair. It results in a ranked opinion lexicon, which is freely available[3].

4.1.2 SO-CAL

[13] constructed their lexicon manually as they believe that the overall accuracy of dictionary-based sentiment analysis mainly relies on the quality of those resources. They built lexicons with content words, namely adjectives, adverbs, nouns and verbs, adding sentiment scores between -5 and +5 (where semantically neutral words are assigned zero score).

4.1.3 SentiWords

SentiWords is a sentiment lexicon derived from SentiWordNet using the method described in [3]. It contains more than 155.000 words associated with a sentiment score between -1 (very negative) and +1 (very positive). The words in this lexicon are arranged with WordNet lists, which include adjectives, nouns, verbs and adverbs.

4.2 The Datasets

In order to evaluate the performance of the proposed lexicons in a sentiment classification task, we used the following two datasets:

4.2.1 Sentiment polarity datasets

This collection[4] consist of 1000 positive and 1000 negative processed reviews. All reviews in this dataset have been extracted from IMDB and Introduced in [11].

4.2.2 Large Movie Review Dataset

This collection of documents[5] reported in [10] consists of 50,000 reviews from IMDB, allowing less than 30 reviews per movie. The dataset consists of two balanced training and test sets, with 25,000 reviews each. The rating scale is larger than in the previous dataset: it goes from 1 to 10. The borderline variable is set to 4, so the negative reviews are assigned values between 1-4, while the positive ones are in the range 7-10.

[3] https://github.com/almatarneh/SPLM-Lexicon
[4] https://www.cs.cornell.edu/people/pabo/movie-review-data/
[5] http://ai.stanford.edu/ amaas/data/sentiment/

4.3 Evaluation

The three lexicons are evaluated on the two datasets of scaled reviews by using the sentiment classifier introduced above in Section 3.2.

Equation 5 is used to compute the f-score F_1, which is the weighted average of the precision, P, and recall, R.

$$F_1 = 2 * \frac{P * R}{P + R} \tag{5}$$

The experimental results are shown in Table 2. By comparing the f-score obtained by the three lexicons, we may conclude that the lexicon we automatically generated, SPLM, consistently outperforms the other manual lexicons on the two datasets.

It is worth noticing that SO-CAL and SentiWords are general-purpose polarity lexicons, while SPLM is a domain-specific resource. This might explain why our lexicon performs better. However, we should point out that SPLM is the result of an automatic method while the other resources were made manually.

Lexicon	**Dataset**	**Negative**			**Positive**		
		Precision	*Recall*	*F*1	*Precision*	*Recall*	*F*1
SPLM	SPD	0.84	0.81	**0.83**	0.81	0.84	**0.83**
	LMRD	0.77	0.75	**0.76**	0.75	0.77	**0.76**
SO-CAL	SPD	0.69	0.69	0.69	0.67	0.67	0.67
	LMRD	0.72	0.68	0.70	0.69	0.73	0.71
SentiWords	SPD	0.72	0.69	0.70	0.69	0.72	0.71
	LMRD	0.69	0.65	0.67	0.67	0.67	0.70

Table 2. Results in terms of precision (P), recall (R), and F_1 scores for Positive and Negative classification. The best F_1 in each datset is highlighted (in bold)

5 Conclusions

Lexicon-based approaches are very popular in sentiment analysis and opinion mining, and they play a key role in all applications in this field. We described in this article a method for automatically building domain-specific polarity lexicons from annotated corpora. A specific lexicon has been built using movie reviews, and we evaluated its quality in an indirect way. More precisely, the lexicon was used to train a sentiment classifier which was evaluated by means of well-known datasets. The experiments reported in our work shows that the lexicon we generated automatically outperforms other manual lexicons when they are used as features of a supervised sentiment classifier. Our corpus-based strategy is not restricted to a particular domain. It is generic enough to be expanded to

whatever domain and language if we are provided with corpora annotated in the appropriate way.

In future work, we will build more domain-specific lexicons for diverse domains in order to compare them again with the general-purpose, and manual lexicons we have used in the present work.

References

1. Augustyniak, L., Kajdanowicz, T., Kazienko, P., Kulisiewicz, M., Tuliglowicz, W.: An approach to sentiment analysis of movie reviews: Lexicon based vs. classification. In: International Conference on Hybrid Artificial Intelligence Systems. pp. 168–178. Springer (2014)
2. 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)
3. 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)
4. Huang, S., Niu, Z., Shi, C.: Automatic construction of domain-specific sentiment lexicon based on constrained label propagation. Knowledge-Based Systems 56, 191–200 (2014)
5. Hutto, C.J., Gilbert, E.: Vader: A parsimonious rule-based model for sentiment analysis of social media text. In: Eighth International AAAI Conference on Weblogs and Social Media (2014)
6. Kamps, J., Marx, M., Mokken, R.J., De Rijke, M., et al.: Using wordnet to measure semantic orientations of adjectives. In: LREC. vol. 4, pp. 1115–1118. Citeseer (2004)
7. Kim, S.M., Hovy, E.: Extracting opinions, opinion holders, and topics expressed in online news media text. In: Proceedings of the Workshop on Sentiment and Subjectivity in Text. pp. 1–8. Association for Computational Linguistics (2006)
8. Lu, Y., Castellanos, M., Dayal, U., Zhai, C.: Automatic construction of a context-aware sentiment lexicon: an optimization approach. In: Proceedings of the 20th international conference on World wide web. pp. 347–356. ACM (2011)
9. Lyu, K., Kim, H.: Sentiment analysis using word polarity of social media. Wireless Personal Communications 89(3), 941–958 (2016)
10. Maas, A.L., Daly, R.E., Pham, P.T., Huang, D., Ng, A.Y., Potts, C.: Learning word vectors for sentiment analysis. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. pp. 142–150. Association for Computational Linguistics, Portland, Oregon, USA (June 2011), http://www.aclweb.org/anthology/P11-1015
11. Pang, B., Lee, L.: A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In: Proceedings of the ACL (2004)
12. Potts, C.: On the negativity of negation. In: Semantics and Linguistic Theory. vol. 20, pp. 636–659 (2010)
13. Taboada, M., Brooke, J., Tofiloski, M., Voll, K., Stede, M.: Lexicon-based methods for sentiment analysis. Computational linguistics 37(2), 267–307 (2011)
14. Turney, P.D.: Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In: Proceedings of the 40th annual meeting on association for computational linguistics. pp. 417–424. Association for Computational Linguistics (2002)