Web Opinion Mining: How to extract opinions from blogs?

Ali Harb, Michel Plantié, Gérard Dray, Mathieu Roche, François Trousset, Pascal Poncelet

Web Opinion Mining: How to extract opinions from blogs?

Ali Harb[*]
EMA-LGI2P
Parc Scientifique G. Besse
30035 Nîmes Cedex, France
ali.harb@ema.fr

Michel Plantié
EMA-LGI2P
Parc Scientifique G. Besse
30035 Nîmes Cedex, France
michel.plantie@ema.fr

Gerard Dray
EMA-LGI2P
Parc Scientifique G. Besse
30035 Nîmes Cedex, France
gerard.dray@ema.fr

Mathieu Roche[†]
LIRMM CNRS 5506
UM II, 161 Rue Ada
F-34392 Montpellier, France
mathieu.roche@lirmm.fr

François Trousset
EMA-LGI2P
Parc Scientifique G. Besse
30035 Nîmes Cedex, France
francois.trousset@ema.fr

Pascal.Poncelet
EMA-LGI2P
Parc Scientifique G. Besse
30035 Nîmes Cedex, France
pascal.poncelet@ema.fr

ABSTRACT

The growing popularity of Web 2.0 provides with increasing numbers of documents expressing opinions on different topics. Recently, new research approaches have been defined in order to automatically extract such opinions from the Internet. They usually consider opinions to be expressed through adjectives, and make extensive use of either general dictionaries or experts to provide the relevant adjectives. Unfortunately, these approaches suffer from the following drawback: in a specific domain, a given adjective may either not exist or have a different meaning from another domain. In this paper, we propose a new approach focusing on two steps. First, we automatically extract a learning dataset for a specific domain from the Internet. Secondly, from this learning set we extract the set of positive and negative adjectives relevant to the domain. The usefulness of our approach was demonstrated by experiments performed on real data.

Categories and Subject Descriptors

H.3.3 [**Information Systems**]: INFORMATION STORAGE AND RETRIEVAL—*Information Search and Retrieval*.

General Terms

Algorithms, Design, Experimentation, Human Factors

Keywords

Text Mining, Opinion Mining, Association Rules, Semantic Orientation.

1. INTRODUCTION

With the fast growing development of the Web, and especially of Web 2.0, there is an ever-increasing number of documents expressing opinions. As an illustration, we shall consider the case of documents giving users' opinions about a camera or a movie, a research topic addressed by various scientific communities including Data Mining, Text Mining and Linguistics. The approaches usually proposed try to identify positive or negative opinion features in order to compile training sets and then apply classification algorithms (based on several linguistic techniques) so as to automatically classify new documents extracted from the Web. They associate semantic opinion orientations with certain adjectives [15, 14, 16, 5, 7]. One of the important issues is thus to define the list of relevant positive and negative adjectives, using either general dictionaries or expert opinions. However, these approaches suffer from the following drawback: in a specific domain, a given adjective may either not exist or have a different meaning from another domain. Consider the following two sentences "*The picture quality of this camera is high*" and "*The ceilings of the building are high*". In the first one, (i.e. an opinion expressed about a camera), the adjective *high* is considered as positive. In the second sentence (i.e. a document on architecture), the same adjective is neutral. This example shows that a given adjective is highly correlated to a particular domain. In the same way, while we may find that *a chair is comfortable*, such an adjective will never be used when talking about movies. In this paper we would like to answer the following two questions: Is it possible to automatically extract a training set from the Web for a particular domain? and How can we extract sets of positive and negative adjectives?

The rest of the paper is organized as follows. Section 2 gives a brief overview of existing approaches for extracting opinions. Our approach, which we call AMOD (*Automatic Mining of Opinion Dictionaries*), is described in Section 3. Section 4 deals with experiments performed on real data sets extracted from blogs. Section 5 sums up the conclusions of the paper.

2. RELATED WORK

[*]Ecole des Mines d'Ales, Laboratoire de Génie Informatique et d'Ingénierie de Production

[†]Laboratoire d'Informatique de Robotique et de Microélectronique de Montpellier, Centre National de la Recherche Scientifique

CSTST 2008 October 27-31, 2008, Cergy-Pontoise, France

As previously mentioned, most approaches consider adjectives as the main source expressing subjective meaning in a given document. Generally speaking, the semantic orientation of a document is determined by the combined effect of the adjectives it contains, on the basis of an annotated dictionary of adjectives labeled as positive or negative (e.g. Inquirer, which contains 3596 words [13] or HM, with 1336 adjectives [5]). More recently, adjective learning have been enhanced by new approaches using such systems as WordNet [8]. These approaches automatically add synonyms and antonyms [2], or extract opinion-related words [16, 6]. The quality of the final result is closely related to the dictionaries available. Moreover, such approaches are not able to detect differences between subject domains (for example the semantic orientation of the adjective "high"). To avoid this problem, more recent approaches use statistical methods based on the co-occurrence of adjectives with an initial set of seed words. The general principle is as follows: beginning with a set of positive and negative words (e.g. *good*, *bad*), to try to extract adjectives situated near to each other according to a measure of distance. The underlying assumption is that a positive adjective appears more frequently next to a positive seed word, and a negative adjective appears more frequently next a negative seed word. While such approaches are efficient, they have the same disadvantages as the previously mentioned techniques with regard to domain-related words.

3. THE AMOD APPROACH

This section presents an overview of the AMOD approach. The general process occurs in three main phases (C.f. figure 1).

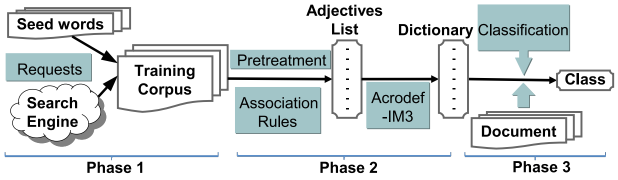

Figure 1: The main process of the AMOD approach

- **Phase 1: Corpora Acquisition Learning Phase.** The aim of this phase is to automatically extract documents containing positive and negative opinions from the Web, for a specific domain.

- **Phase 2: Adjective Extraction Phase.** In this phase, we automatically extract sets of relevant positive and negative adjectives.

- **Phase 3: Classification.** The aim of this phase is to classify new documents using the sets of adjectives obtained in the previous phase.

In this paper, we particularly focus on the first two phases. The classification task currently uses very simple operations and will be enhanced by future research work.

3.1 Phase 1: Corpora Acquisition Learning Phase

In order to identify relevant adjectives, we first focus on the automatic extraction of a training set for a specific domain. We therefore consider 2 sets P and N of seed words with positive and negative semantic orientations respectively, as in [15].

$$P = \{ \quad \textit{good}, \textit{nice}, \textit{excellent}, \textit{positive}, \textit{fortunate}, \textit{correct}, \textit{superior} \}$$

$$N = \{ \quad \textit{bad}, \textit{nasty}, \textit{poor}, \textit{negative}, \textit{unfortunate}, \textit{wrong}, \textit{inferior} \}$$

For each seed word, we use a search engine, entering a request specifying: the application domain d, the seed word we are looking for, and the words we want to avoid. For example, using the search engine Google, which specializes in blog searching, in order to obtain opinions about movies containing the seed word *good*, the following request is sent "+opinion +review +movies +*good* -bad -nasty -poor -negative -unfortunate -wrong -inferior". The results given by this request will be documents giving opinions about cinema containing the word *good* and without the following words: bad, nasty, poor, ... inferior. Therefore, for each positive (resp. negative) seed word in a given domain, we automatically collect K documents in which none of the negative (resp. positive) adjective set appears. This operation generates 14 learning corpora: 7 positive and 7 negative.

3.2 Phase 2: Adjective Extraction Phase

The corpora built up in the previous phase provide us with documents containing domain-relevant seed adjectives. This phase therefore focuses on extracting adjectives from those domain-relevant documents that are highly correlated with the seed adjectives. Within the collected document corpora, we compute correlations between the seed words and other adjectives so as to enrich the sets of seed words with new domain-relevant opinion adjectives. However, in order to avoid false positive or false negative adjectives we also add new filtering steps. These steps are presented in the following subsections.

3.2.1 Preprocessing and Association Rules Steps

In order to compute the correlations between adjectives that will enrich the opinion dictionary, we must determine the Part-of-Speech tag (Verb, Noun, Adjective, etc.) of each word in the learning corpus. To do this, we use the tool Tree Tagger [12], which automatically gives a Part-of-Speech tag for each word in a text and converts it to its lemmatized form. As in [14, 16, 5, 7], we consider adjectives as representative words for specifying opinions. From the TreeTagger results we therefore only retain the adjectives embedded in the documents. We then search for associations between the adjectives contained in the documents and the seed words in the positive and negative seed sets. The aim is to find out whether any new adjectives are associated with the same opinion polarity as the seed words. In order to obtain these correlations, we adapted an association rule algorithm [1] to our purposes. More formally, let $I = \{adj_1, ...adj_n\}$ be a set of adjectives, and D a set of sentences, where each sentence corresponds to a subset of elements of I. An association rule is thus defined such that X→Y, where X$\subset I$, Y$\subset I$, and X $\cap$ Y = $\oslash$. The support of this rule corresponds to the percentage of sentences in D containing X∪Y. The rule X→Y has a confidence ratio c, if c% of sentences from D containing X also contain Y.

A sentence could be considered as a part of text separated off by particular punctuation marks. However, in order to obtain more relevant adjectives we applied the following hypothesis: the closer a given adjective is to a seed adjective, the greater the similarity in semantic orientation between the two adjectives. We thus define sentences in terms of window size (WS). The WS is the distance between a seed word and an adjective. For instance, if the WS is set to 1 that means that a sentence is made up of one adjective before the seed word, the seed word itself and one adjective after it. In the following sentence, "*The movie is amazing, good acting, a lot of great action and the popcorn was delicious*", by considering the seed adjective *good*, with WS=1 we obtain the following sentence: "*amazing, good, great*", and with WS=2: "*amazing, good, great, delicious*".

The association rule algorithm is applied to both the positive and the negative corpora. At the end of this step, we are thus provided with a number of rules about adjectives. An example of such a rule is: *amazing, good* → *funny* meaning that when a sentence contains *amazing* and *good*, then very often (in function of a support value s) we can also find *funny*.

3.2.2 Filtering step

As we are interested in adjectives that are strongly correlated with the seed words, from the results obtained in the previous step we only retain rules containing more than one seed word. We then consider adjectives appearing in both the positive and the negative list. Those correlated to more than one seed word with the same orientation and a high support value are retained as learned adjectives only if the number of times they occur in each document of one corpus (e.g. the positive one) is greater than 1 and the number of occurrences in each document in the other corpus (e.g. the negative one) is lower than 1. Otherwise they are removed. Finally, in order to filter the associations extracted in the previous step, we use a ranking function to generate a list and delete the irrelevant adjective associations at the end of the list. One of the most commonly used measures for finding how two words are correlated (i.e. is there a co-occurrence relationship between the two words) is Cubic Mutual Information ($MI3$) [4]. This empirical measure, based on Church's Mutual Information (MI) [3], enhances the impact of frequent co-occurrences. Our approach relies on computing the dependence of two adjectives based on the number of pages on the Web returned by the queries "$adjective_1$ $adjective_2$" and "$adjective_2$ $adjective_1$"[1]. This dependence is computed in a given context C (e.g. the context $C = \{movies\}$). Then we apply the formula $AcroDef_{MI3}$ (1) described in [11].

$$AcroDef_{MI3}(adj1, adj2) =$$

$$\frac{(nb(\text{"}adj1\ adj2\text{"}\ and\ C) + nb(\text{"}adj2\ adj1\text{"}\ and\ C))^3}{nb(adj1\ and\ C) \times nb(adj2\ and\ C)} \quad (1)$$

3.3 Phase 3: Classification

The last phase to consider is the classification of each document according to positive or negative opinions. In the first step, we use a very simple classification procedure. For each document to be classified, we calculate its positive or negative orientation by computing the difference between the number of positive and negative adjectives encountered, from both of the previously described lists. We count the number of positive adjectives, then the number of negative adjectives, and simply compute the difference. If the result is positive (resp. negative), the document will be classified in the positive (resp. negative) category. Otherwise, the document is considered to be neutral.

In order to improve the classification, we extend our method to consider any adverbs or other words used to invert the polarities (e.g. not, neither, nor, etc.) For instance, if we consider the following sentence: *The movie is not bad, there are a lot of funny moments*. The adverb *not* inverses the polarity of the adjective *bad* while the word *funny*, being too far from the word *not*, is not affected. Furthermore, for the following adverbs: very, so, too, we increase the degree of semantic orientation by 30%. This figure 30% is chosen to increase the semantic orientation of the considered adjectives in a rather moderate proportion. In the future this value may change according to experimental results.

4. EXPERIMENTS

This section describes the experiments we carried out in order to validate our approach. First we present the results of the adjective learning and classification phases, then we compare our method to a supervised machine learning classification method.

The documents were extracted from the research engine BlogGooglesearch.com. We extracted documents related to opinions expressed in the domain of "cinema". The seed words and requests applied were those already mentioned in Section 3.1. For each seed word, we limited the number of documents extracted by the search engine to 300. We then transformed these documents from HTML format to text format and used TreeTagger to retain only the adjectives.

In order to study the best distance between seed words and adjectives to be learned, we tested different values for the Window Size parameter from 1 to 3. Then, to extract correlation links between adjectives, we use the Apriori algorithm[2]. In the experiments conducted, the support value ranged from 1 to 3%. For each support value, we obtained two lists: one negative and one positive. As described above, we discarded from these lists any adjectives that were common to both lists (for the same support value) and those which correlated to only one seed word. To discard useless and frequent adjectives we used the AcroDef$_{MI3}$ measure with a threshold value experimentally fixed at 0.005.

In order to test the quality of the adjectives learned, for classification purposes we used the Movie Review Data from the NLP Group, Cornell University[3]. This database contains 1000 positive and 1000 negative opinions extracted from the Internet Movie Database[4]. We intentionally used a test corpus very different in nature from the training corpora (i.e. blogs), in order to show the stability of our method. Table 1 shows the classification results considering only the seed words (i.e. without applying the AMOD approach) for the negative and positive corpora. PL (resp. NL) corre-

[1] Here we assume that the request is made using Google, so the brackets stand for the real string respecting the order between adjectives

[2] http://fimi.cs.helsinki.fi/fimi03/

[3] http://www.cs.cornell.edu/people/pabo/movie-review-data/

[4] http://www.imdb.com/

	Positives	Negatives	PL	NL
Seed List	**66,9%**	**30,4%**	7	7

Table 1: Classification of 1000 positive and negative documents with seed words

sponds to the number of adjectives (in our case, this number corresponds to the number of seed words).

WS	S	Positive	PL	NL
1	1%	**67,2%**	7+12	7+20
	2%	60,3%	7+8	7+13
	3%	65,6%	7+6	7+1
2	1%	57,6%	7+13	7+35
	2%	56,8%	7+8	7+17
	3%	68,4%	7+4	7+4
3	1%	28,9%	7+11	7+48
	2%	59,3%	7+4	7+22
	3%	67,3%	7+5	7+11

Table 2: Classification of 1000 positive documents with learned adjectives

WS	S	Negative	PL	NL
1	1%	**39,2%**	7+12	7+20
	2%	46,5%	7+8	7+13
	3%	17,7%	7+6	7+1
2	1%	49,2%	7+13	7+35
	2%	49,8%	7+8	7+17
	3%	32,3%	7+4	7+4
3	1%	76,0%	7+11	7+48
	2%	46,7%	7+4	7+22
	3%	40,1%	7+5	7+11

Table 3: Classification of 1000 negative documents with learned adjectives

Table 2 (resp. Table 3), shows the results obtained with learned adjectives using Amod after classifying the positive (resp. negative) documents. Column WS corresponds to the distances and column S to the support values. The value 7 + 12 in the first line of the PL column indicates that we have 7 seed adjectives and 12 learned adjectives. As you can see, our method enables much better classification results for negative documents. In the case of positive documents, the difference is smaller but, as illustrated in Table 4, the learned adjectives appear with very significant frequency in the test documents.

As expected if we compare the number of learned adjectives, the best results come with the WS value 1. This experiment confirmed our hypothesis concerning adjective proximity in expressions of opinion [15]. In table 2 and 3, we see that numbers of positive and negative learned adjectives may vary considerably in function of the support value. For example, if the support value is 1% and WS=3, we obtain 11 positive learned adjectives and 48 negative ones. Thorough analysis of the results showed that most of the negative adjectives were frequent and useless adjectives. The results obtained by applying the AcroDef$_{MI3}$ measure as an adjective filter are given in Tables 6 and 7, which only includes results obtained with WS=1 and S=1%. The proportion of documents that were well classified by our approach ranges from 66.9% to 75.9% for positive adjectives and from 30.4% to 57.1% for negative adjectives. To enhance our method and extract the best discriminative adjectives, we applied the following method:

positive seeds		negative seeds	
Adjective	Nb of occ.	Adjectives	Nb of occ.
Good	2147	Bad	1413
Nice	184	Wrong	212
Excellent	146	Poor	152
Superior	37	Nasty	38
Positive	29	Unfortunate	25
Correct	27	Negative	22
Fortunate	7	Inferior	10

Table 4: Occurrences of positive and negative seed adjectives for WS=1 and S=1%

Learned positive adjectives			
Adjective	Nb of occ.	Adjective	Nb of occ.
Great	882	Hilarious	146
Funny	441	Happy	130
Perfect	244	Important	130
Beautiful	197	Amazing	117
Worth	164	Complete	101
Major	163	Helpful	52

Table 5: Occurrences of positive learned adjectives for WS=1 and S=1%

- We enriched the seed word lists with adjectives learned from the previous application of Amod in order to obtain new seed word lists.
- We then applied the Amod approach to the new lists so as to learn new adjectives.
- In order to evaluate the new lists, we then applied the classification procedure to the test dataset.

This method was repeated until no more new adjectives were learned. The adjectives learned by applying this reinforcement method for the first time are shown in Table 8. Our adjective set was thus enriched by the learned adjectives considered to be relevant and representative. The results obtained in the classification step are shown in Table 9. The ratio of correctly attributed positive documents was improved by the second reinforcement learning phase, going from 75.9 to **78.1%**.

The adjectives learned by means of the first reinforcement are then added to the previous seed word lists and the pro-

Learned negative adjectives			
Adjectives	Nb of occ.	Adjectives	Nb of occ.
Boring	200	Certain	88
Different	146	Dirty	33
Ridiculous	117	Social	33
Dull	113	Favorite	29
Silly	97	Huge	27
Expensive	95		

Table 6: Occurrences of negative learned adjectives for pour WS=1 et S=1%

WS	S	Positive	Negative	PL	NL
1	1%	**75,9%**	**57,1%**	7+11	7+11

Table 7: Classification of 1000 positive and negative documents with learned adjectives and AcroDef$_{MI3}$

Learned positive adj.	
Adjectives	Nb of occ.
Interesting	301
comic	215
Wonderful	165
Successful	105
Exciting	88

Learned negative adj.	
Adjectives	Nb of occ.
Commercial	198
Dead	181
Terrible	113
Scary	110
Sick	40

Table 8: Learned adjective occurrences with the first reinforcement for WS=1 and S=1%

WS	S	Positive	Negative	PL	NL
1	1%	**78,1%**	**54,9%**	7+16	7+16

Table 9: Classification of 1000 positive and negative documents with learned adjectives and AcroDef$_{MI3}$

cess is repeated. The second reinforcement phase produces new adjectives (C.f. Table 10).

Learned positive adj.	
Adjectives	Nb of occ.
special	282
entertaining	262
sweet	120

Learned negative adj.	
Adjectives	Nb of occ.
awful	109

Table 10: Learned adjective occurrences with the second reinforcement for WS=1 et S=1%

Table 11 shows that the results of the classification of positive documents improved from 78.1% to **78.7**%, for the same dataset test. However, the results are slightly lower for negative documents. This can be explained by the excessively elementary nature of the classification procedure, based on the number of adjective occurrences. The list of learned adjective shows that the number of occurrences of positive learned adjectives is notably greater than those of learned negative adjectives. This significantly influences our classification results.

WS	S	Positive	Negative	PL	NL
1	1%	**78,7%**	**46,7%**	7+16	7+16

Table 11: Classification of 1000 positive and negative documents with learned adjectives and AcroDef$_{MI3}$

We improved our classification method by adding the various forms of negation presented in previous section. Our results for the classification of 1000 positive texts improved from 78.7% to **82.6%** and from 46.7% to **52.4%** for the 1000 negative texts as shown in Table 12.

Further re-application of the reinforcement learning phase did not produce any new adjectives. At the end of the process we obtained two relevant and discriminatory adjective lists (C.f. Table 13) for the *cinema* domain.

WS	S	Positive	Negative	PL	NL
1	1%	**82,6%**	**52,4%**	7+19	7+17

Table 12: Classification of 1000 positive and negative documents classification with learned adjectives, AcroDef$_{MI3}$ and negation

Positive adjective list	
Adjective	Adjective
Good	Great
Nice	Funny
Excellent	Perfect
Superior	Beautiful
Positive	Worth
Correct	Major
Fortunate	Interesting
Hilarious	Comic
Happy	Wonderful
Important	Successful
Amazing	Exciting
Complete	Entertaining
Special	Sweet

Negative adjective list	
Adjective	Adjective
Bad	Boring
Wrong	Different
Poor	Ridiculous
Nasty	Dull
Unfortunate	Silly
Negative	Expensive
Inferior	Huge
Certain	Dead
Dirty	Terrible
Social	Scary
Favorite	Sick
Awful	Commercial

Table 13: Adjective lists for WS=1 and S=1% for the domain "*cinema*"

In this experiment, we wanted to find out how many documents were required to produce a stable and robust training set? We therefore applied the Amod training method several times. We increased the number of collected documents by 50 each time until the number of learned adjectives remained stable.

Figure 2 depicts the relationship between the size of the corpus and the number of learned adjectives. As we can observe, above 2800 documents (i.e. 200 documents for each seed word) not many new adjectives are learned.

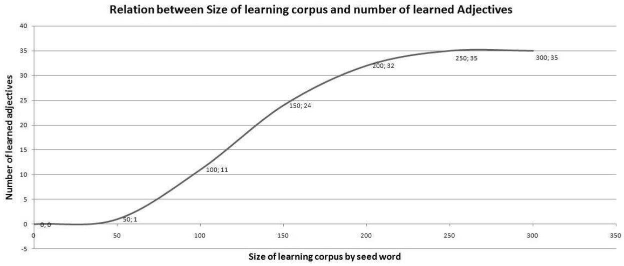

Figure 2: Relation between the size of training corpus and the number of learned adjectives

Finally we conducted some experiments in order to compare the results obtained using a traditional classification method with our approach. The classification method used for the experiments was Copivote [9]. This approach use a training corpus and a system of vote with several classifiers (SVM, ngrams, ...). Experiments were performed on the same datasets for learning and tests.

To compare our results, we used the well known FScore measure [10]. FScore is given by the following formula:

$$Fscore = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

Fscore is a compound between Recall and Precision, giving the same weight to each measure. *Precision* and *Recall* are defined as follows:

$$Recall_i = \frac{Nb\ documents\ rightly\ attributed\ to\ class\ i}{Nb\ documents\ of\ class\ i}$$

$$Precision_i = \frac{Nb\ documents\ rightly\ attributed\ to\ class\ i}{Nb\ documents\ attributed\ to\ class\ i}$$

Documents :	Positives	Negatives
FScore COPIVOTE :	60,5%	60,9%
FScore AMOD :	**71,73%**	**62,2%**

Table 14: Fscore classification results for 1000 negative and positive test documents with COPIVOTE and AMOD

Table 14 shows that our approach performs better for both the positive (**71,73%** vs. 60,5%) and the negative case (**62,2%** vs. 60,9%). Generally the COPIVOTE method is very efficient for text classification (i.e. based on a voting system, the best classification method is selected), but is penalized by the large differences between the test and training corpora.

In order to verify that our approach is suitable for other domains we performed some experiments with a totally different domain: "car". Positive and Negative corpora were obtained using BlogGooglesearch.com with the keyword "*car*". To validate knowledge acquired in the training phase, we used 40 positive documents in the test phase, derived from *www.epinions.com.*
Applying the AMOD approach, with WS=1 and support = 1%, plus AcroDef$_{IM3}$ filtering and reinforcement training gave the results shown in Table 15.
We optained the following positive adjectives: good, nice, excellent, superior, positive, correct, fortunate, professional, popular, luxurious, secured, great, full, efficient, hard, fast, comfortable, powerful, fabulous, economical, quiet, strong, several, lovely, successful, amazing, maximum, first, active, beautiful, wonderful, practical.
And the following negative adjectives: bad, wrong, poor, nasty, unfortunate, negative, inferior, horrible, boring, unsecured, uncomfortable, expensive, ugly, luck, heavy, dangerous, weird.

Method	WS	S	Positive	PL	NL
Seed words only	1	1%	**57,5%**	7+0	7+0
with learned words	1	1%	**95%**	7+26	7+10

Table 15: 40 positive documents Classification with seed adjectives only and with learned adjectives, AcroDef$_{IM3}$ and negation filters

Compared to previous experiments, both training sets were similarly constituted from blogs. Our approach gave better results for similar data sets.

5. CONCLUSION

In this paper, we proposed a new approach for automatically extracting positive and negative adjectives in the context of opinion mining. Experiments conducted on training sets (blogs vs. cinema reviews) showed that our approach was able to extract relevant adjectives for a specific domain. There is a great deal of scope for future work. Firstly, our method depends on the quality of the documents extracted from blogs. We want to extend our training corpora method by applying text mining approaches to collected documents in order to minimize lower-quality, noisy texts. Secondly, while in this paper we focused on adjectives, we plan to extend the extraction task to other word categories.

6. REFERENCES

[1] R. Agrawal and R. Srikant. Fast algorithms for mining association rules in large databases. In *VLDB'94*, 1994.

[2] A. Andreevskaia and S. Bergler. Semantic tag extraction from wordnet glosses. 2007.

[3] K. Church and P. Hanks. Word association norms, mutual information, and lexicography. In *Computational Linguistics*, volume 16, pages 22–29, 1990.

[4] D. Downey, M. Broadhead, and O. Etzioni. Locating complex named entities in web text. In *Proceedings of IJCAI'07*, pages 2733–2739, 2007.

[5] V. Hatzivassiloglou and K. McKeown. Predicting the semantic orientation of adjectives. In *Proceedings of 35th Meeting of the Association for Computational Linguistics*, Madrid, Spain, 1997.

[6] M. Hu and B. Liu. Mining and summarizing customer reviews. In *Proceedings of KDD'04, ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Seattle, WA, 2004.

[7] J. Kamps, M. Marx, R. J. Mokken, and M. Rijke. Using wordnet to measure semantic orientation of adjectives. In *Proceedings of LREC 2004, the 4th International Conference on Language Resources and Evaluation*, pages 174–181, Lisbon, Portugal, 2004.

[8] G. Miller. Wordnet: A lexical database for english. In *Communications of the ACM*, 1995.

[9] M. Plantié, M. Roche, G. Dray, and P. Poncelet. Is a voting approach accurate for opinion mining? In *Proceedings of the 10th International Conference on Data Warehousing and Knowledge Discovery (DaWaK '08), Torino Italy*, 2008.

[10] V. Risbergen. Information retrieval, 2nd edition. In *Butterworths*, London, 1979.

[11] M. Roche and V. Prince. *AcroDef:* A Quality Measure for Discriminating Expansions of Ambiguous Acronyms. In *Proceedings of CONTEXT, Springer-Verlag, LNCS*, pages 411–424, 2007.

[12] H. Schmid. Treetagger. In *TC project at the Institute for Computational Linguistics of the University of Stuttgart*, 1994.

[13] P. Stone, D. Dunphy, M. Smith, and D. Ogilvie. The general inquirer: A computer approach to content analysis. Cambridge, MA, 1966. MIT Press.

[14] M. Taboada, C. Anthony, and K. Voll. Creating semantic orientation dictionaries. 2006.

[15] P. Turney. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In *Proceedings of 40th Meeting of the Association for Computational Linguistics*, pages 417–424, Paris, 2002.

[16] K. Voll and M. Taboada. Not all words are created equal: Extracting semantic orientation as a function of adjective relevance. pages 337–346. Volume 4830/2007 AI 2007: Advances in Artificial Intelligence, 2007.