DICTIONARY EXTENSION FOR IMPROVING AUTOMATED SENTIMENT DETECTION

Johannes Liegl, Stefan Gindl, Arno Scharl and Alexander Hubmann-Haidvogel
Department of New Media Technology, MODUL University Vienna, Am Kahlenberg 1, Vienna, Austria

Keywords: Sentiment detection, Natural language processing, Latent semantic analysis, Pointwise mutual information.

Abstract: This paper investigates approaches to improve the accuracy of automated sentiment detection in textual knowledge repositories. Many high-throughput sentiment detection algorithms rely on sentiment dictionaries containing terms classified as either positive or negative. To obtain accurate and comprehensive sentiment dictionaries, we merge existing resources into a single dictionary and extend this dictionary by means of semi-supervised learning algorithms such as Pointwise Mutual Information - Information Retrieval (PMI-IR) and Latent Semantic Analysis (LSA). The resulting extended dictionary is then evaluated on various datasets from different domains, which were annotated on both the document and sentence level.

1 INTRODUCTION

In recent years the field of opinion mining and sentiment analysis has attracted a lot of attention from the data mining research community. The research community noticed the vast and ever growing amount of freely available user-generated data such as product and movie reviews on the web, as well as the labels and ratings the users provided to precisely sum up their opinions. Under these ideal circumstances it is not surprising that the application and tailoring of text mining, machine learning and information retrieval techniques towards the problem of sentiment classification has been examined so intensively.

In this paper we describe a way to extend our original sentiment dictionary by merging it with another one and searching for similar terms in a web-based news corpus and propagating the sentiment values of the original terms to the newly identified ones. To search for similar terms we use Latent Semantic Analysis (LSA) and Pointwise Mutual Information - Information Retrieval (PMI-IR). Our evaluation shows that this approach leads to gains in recall that outperform losses in precision over all of our test datasets.

In Section 2 we describe related work in the field of sentiment detection and differentiate it from our own. Section 3 describes how we used LSA and PMI-IR to find similar values and how the propagation of the sentiment values works. Section 4 presents an evaluation of extended dictionaries on various datasets and Section 5 concludes the paper.

2 RELATED WORK

The current research in sentiment detection can broadly be divided into machine learning and knowledge-based approaches. The ones backed by machine learning techniques try to learn a model from labeled training data and predict the labels of test data. (Pang et al., 2002) train Naïve Bayes (NB), Support Vector Machine (SVM) and Maximum Entropy (ME) classifiers on movie reviews using various combinations of unigram, bigram and part-of-speech features. (Mullen and Collier, 2004) train SVMs on movie reviews and incorporate features based on pointwise mutual information and the emotive meaning of adjectives from WordNet. In our previous research (Gindl and Liegl, 2008) we could reproduce the findings of (Pang et al., 2002) and achieve similar accuracy values for ME-based classifiers on datasets from the travel and consumer products domain.

The sentiment detection approaches in the knowledge-based strand of research use seed-terms with known sentiment values (sentiment dictionaries) in various ways to perform sentiment classification. (Turney, 2002) uses the seed-terms ”poor” and ”excellent” to calculate the mutual information between these terms and phrases from the text to classify. (Yu and Hatzivassiloglou, 2003) employ a co-occurrence measure on a set of seed-terms and a large news-wire corpus to find new sentiment carrying terms and their associated sentiment value. In (Read and Carroll, 2009) different word similarity measures like lexi-

Liegl J., Gindl S., Scharl A. and Hubmann-Haidvogel A..
DICTIONARY EXTENSION FOR IMPROVING AUTOMATED SENTIMENT DETECTION.
DOI: 10.5220/0003070304040407
In *Proceedings of the International Conference on Knowledge Discovery and Information Retrieval* (KDIR-2010), pages 404-407
ISBN: 978-989-8425-28-7

cal association (i.e. PMI-IR), semantic spaces (i.e. LSA) and distributional similarity are applied on a small seed set of 14 terms to determine the sentiment value of every term in the testing data. Their approach does not achieve as good results as supervised learning methods but it is independent from the availability of training data and largely domain-independent.

Our approach also falls into the knowledge-based category, and we also use PMI-IR and LSA for dictionary extension. We also try to combine the results of PMI-IR and LSA into a hybrid dictionary. Additionally, we also differentiate between datasets that are annotated on the sentence and document level.

3 METHODOLOGY

Two methods to extend sentiment dictionaries are examined in the following. Subsection 3.1 describes the used dictionaries, Subsection 3.2 outlines the simple merging procedure. Subsections 3.3 and 3.4 discuss dictionary extension using LSA and PMI-IR.

3.1 Sentiment Dictionaries

A sentiment dictionary (*SD*) is a listing of opinionated terms, where each term has assigned a value from $[-1,1]$. Negative values denote negative terms and vice versa. We used the following three dictionaries:

General Inquirer (*GI*). After eliminating neutral and ambiguous terms from the originally 11 789 words created by (Stone et al., 1966) we ended up with a list of 3 682 words.

Subjectivity Lexicon (*SL*). This lexicon is used by (Wilson et al., 2005). Filtering insecure terms (e.g. ambiguous sentiment polarity) from the originally 8 221 terms we ended up with a list of 4 755 terms.

Semantic Word List (*SW*). The Semantic Word List is a *SD* created and used in previous projects at our institute. It is based on the *GI* and was augmented with terms from political blog posts. The list contains 8 276 terms, and will be the benchmark during evaluation.

3.2 Dictionary Merging

The most intuitive and easiest way to extend a *SD* is to merge it with another one. In our case we merged *GI* with *SL* in two different ways:

- *GIandSL* is the intersection of *GI* and *SL*, where terms with inconsistent sentiment values have been discarded. This merging leads to a *SD* that contains 2 030 terms.
- *GIorSL* contains all terms in both *GI* and *SL*. Again, terms occurring in both dictionaries with differing sentiment values are discarded. This merging leads to a *SD* that contains 6 407 terms.

3.3 Similar Terms Identification

We used two different techniques to identify similar terms:

Pointwis Mutual Information (PMI-IR) uses Pointwise Mutual Information to measure the likeliness of co-occurrence of two terms (Turney, 2001). A high value indicates a semantic relatedness of the two terms. We use the following formula:

$$score(t,ct_i) = \log_2\left(\frac{p(t\ \&\ ct_i)}{p(t)p(ct_i)}\right) \quad (1)$$

where $p(t\&ct_i)$ is the probability of a fixed term t co-occurring with a candidate term ct_i of an arbitrary corpus.

Latent Semantic Analysis (LSA) uses a dimensionality reduction technique called Singular Value Decomposition (SVD) to analyze the relations among terms in a corpus (Landauer and Dumais, 1997; Deerwester et al., 1990), and is able to discover relations between terms occurring in different documents. Simply spoken, although term *A* and *C* do not co-occur in the same document they can have a relation via term *B*. LSA is capable of discovering such hidden relations. We used the tool JLSI[1] to accomplish LSA.

A corpus of 100 000 documents (crawled from news sites over the period of one year and later on referred to as Media Corpus, *MC*) served as input for LSA and PMI-IR.

3.4 Sentiment Propagation

Sentiment propagation consists of two steps: (i) creation of similar term lists, and (ii) propagating sentiment values to other terms. Figure 1 shows the creation of the similar terms list. The input to the similarity method, which can either be PMI-IR or LSA, is a corpus and an *SD*. The corpus we used for our experiments is the *MC* and the sentiment dictionary is *GIorSL*. For each term $s_i \in SD$ a list sl_i is created that contains all the terms $t_k \in MC$ and the similarity values sv_{ik} that were calculated for each (s_i,t_k) pair

[1] http://tcc.itc.it/research/textec/tools-resources/jlsi.html

by the similarity method. These lists are normalized and sorted by decreasing similarity values. As Figure 1 shows, the same term t_1 will show up on different positions in the lists reflecting the fact that t_1 has different similarity values with different terms from the *SD* (dotted line).

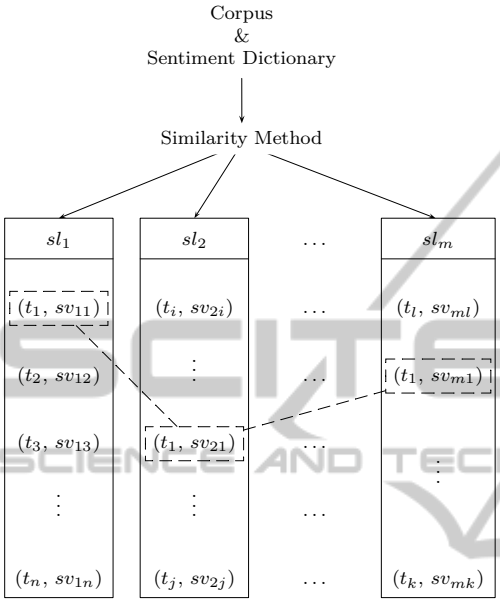

Figure 1: Creation of lists containing similar terms for all terms in a *SD*.

Once the similar term lists have been created the sentiment propagation takes place. Each term $t_i \in MC$ gets its sentiment value s_i using the following formula:

$$s_i = \sum_{(t_k,s_k)\in SD} s_k sv_{ki} \quad (2)$$

The normalized similarity values for term t_i are just multiplied by the original sentiment values and summed up. The s_is are normalized too. Figure 2 shows the sentiment propagation for the term *accuse*. It gets a sentimental charge from each of the terms in the *SD* whose strength is determined by the similarity of *accuse* to the terms in the *SD*. The underlying idea here is that if a newly identified term is more similar to more positive terms from the *SD* than to negative ones it gets a positive sentiment value; otherwise, its value is negative.

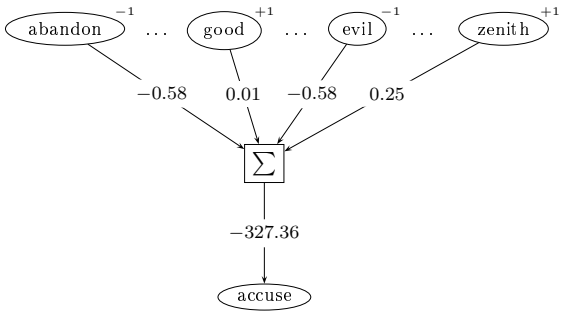

Figure 2: Propagation of sentiment values to a newly identified similar term.

4 EVALUATION

For evaluation and LSA parameter tuning, we used datasets with equal number of positive and negative documents from three different domains: vacation reviews[2], customer reviews[3], and sentences containing a political statement obtained by using a symmetric verification game (Rafelsberger and Scharl, 2009) on the Facebook [4] social-networking platform.

After merging and extending the sentiment dictionaries we added three more dictionaries to *SW* and *GIorSL* (explained in Subsection 3.1 and 3.2, respectively):

- The *GIorSL* sentiment dictionary extended by 4 000 terms using LSA as similarity method.
- The *GIorSL* sentiment dictionary extended by 4 000 terms using PMI-IR as similarity method.
- The *GIorSL* sentiment dictionary extended by 4 000 terms using LSA and 4.000 terms using PMI-IR as similarity method.

Extending a dictionary using one of the described similarity methods means that we added the 2 000 highest ranking (most positive) and the 2 000 lowest ranking (most negative) out of 250 000 terms.

We achieved the best results with the hybrid method using both LSA and PMI, each method contributing 4 000 terms (also see Figure 3). As the improvements for all used combinations were rather low, we decided to aggregate the recall and precision values into one measurement, which we called Precision Recall Gain (PRG). PRG expresses the gain (sum) of the difference between two *SDs* for precision and the difference between two *SDs* for recall. The other used combinations only delivered very ambiguous results, where an increase of recall for positive reviews led to a decrease of recall for negative reviews, or where a gain in recall worsened precision.

[2]www.tripadvisor.com
[3]www.amazon.com
[4]www.facebook.com

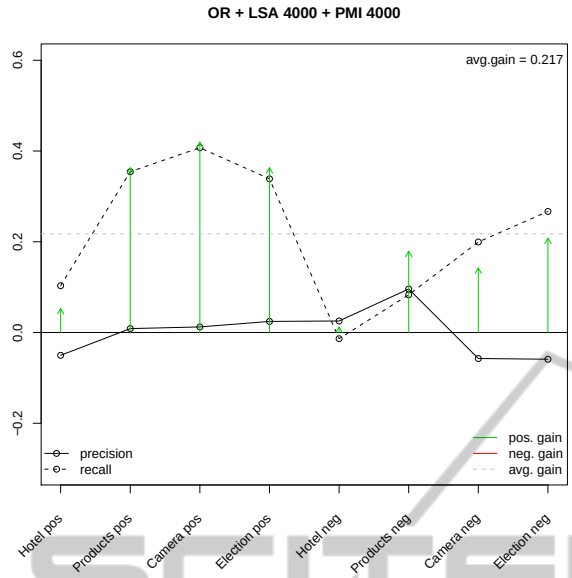

Figure 3: Evaluation results of the different merging strategies. The arrows indicate a gain or loss in PRG.

5 CONCLUSIONS

The evaluation shows small improvements using only LSA or a combination of LSA and PMI-IR. Yet, we believe that their application to smaller dictionaries would have more effect. Using a combined measurement like the proposed Precision-Recall-Gain helps highlighting small improvements. As a future work we see the exploration of different levels of document granularity. Using sentences or paragraphs as the unit for indexing could improve the proposed extension strategies.

ACKNOWLEDGEMENTS

The RAVEN Research Project (Relation Analysis and Visualization for Evolving Networks; www.modul.ac.at/nmt/raven) is funded by the Austrian Ministry of Transport, Innovation & Technology (BMVIT) and the Austrian Research Promotion Agency (FFG) within the strategic objective FIT-IT Semantic Systems (www.fit-it.at). We would like to thank Albert Weichselbraun for providing the data sets for the evaluation and Gerhard Wohlgenannt for the PMI-IR Analysis. We also thank Jode Ziegenfuß for proof reading the manuscript.

REFERENCES

Deerwester, S. C., Dumais, S. T., Landauer, T. K., Furnas, G. W., and Harshman, R. A. (1990). Indexing by latent semantic analysis. *Journal of the American Society of Information Science*, 41(6):391–407.

Gindl, S. and Liegl, J. (2008). Evaluation of different sentiment detection methods for polarity classification on web-based reviews. In *Proceedings of the ECAI Workshop on Computational Aspects of Affectual and Emotional Interaction*.

Landauer, T. K. and Dumais, S. T. (1997). A solution to plato's problem: The latent semantic analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review*, 104:211–240.

Mullen, T. and Collier, N. (2004). Sentiment analysis using support vector machines with diverse information sources. In Lin, D. and Wu, D., editors, *Proceedings of EMNLP 2004*, pages 412–418, Barcelona, Spain. Association for Computational Linguistics.

Pang, B., Lee, L., and 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)*.

Rafelsberger, W. and Scharl, A. (2009). Games with a purpose for social networking platforms. In *Proceedings of the 21st ACM Conference on Hypertext and Hypermedia*.

Read, J. and Carroll, J. (2009). Weakly supervised techniques for domain-independent sentiment classification. In *Proceedings of the 1st International CIKM Workshop on Topic-Sentiment Analysis for Mass Opinion Measurement*.

Stone, P. J., Dunphy, D. C., and Smith, M. S. (1966). *The General Inquirer : A Computer Approach to Content Analysis*. MIT. Press, Cambridge, Mass. [u.a.].

Turney, P. (2002). Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*.

Turney, P. D. (2001). Mining the web for synonyms: PMI–IR versus LSA on TOEFL. In *Proceedings of the 12th European Conference on Machine Learning*.

Wilson, T., Wiebe, J., and Hoffmann, P. (2005). Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of Human Language Technologies Conference/Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP 2005)*, Vancouver, CA.

Yu, H. and 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*, pages 129–136, Morristown, NJ, USA. Association for Computational Linguistics.