Unsupervised Classification of Sentiment and Objectivity in Chinese Text

Taras Zagibalov John Carroll

University of Sussex
Department of Informatics
Brighton BN1 9QH, UK

{T.Zagibalov,J.A.Carroll}@sussex.ac.uk

Abstract

We address the problem of sentiment and objectivity classification of product reviews in Chinese. Our approach is distinctive in that it treats both positive / negative sentiment and subjectivity / objectivity not as distinct classes but rather as a continuum; we argue that this is desirable from the perspective of would-be customers who read the reviews. We use novel unsupervised techniques, including a one-word 'seed' vocabulary and iterative retraining for sentiment processing, and a criterion of 'sentiment density' for determining the extent to which a document is opinionated. The classifier achieves up to 87% F-measure for sentiment polarity detection.

1 Introduction

Automatic classification of sentiment has been a focus of a number of recent research efforts (e.g. (Turney, 2002; Pang et al., 2002; Dave at al., 2003). An important potential application of such work is in business intelligence: brands and company image are valuable property, so organizations want to know how they are viewed by the media (what the 'spin' is on news stories, and editorials), business analysts (as expressed in stock market reports), customers (for example on product review sites) and their own employees. Another important application is to help people find out others' views about products they have purchased (e.g. consumer electronics), services and entertainment (e.g. movies), stocks and shares (from investor bulletin boards), and so on. In the work reported in this paper we focus on product reviews, with the intended users of the processing being would-be customers.

Our approach is based on the insight that positive and negative sentiments are extreme points in a *continuum* of sentiment, and that intermediate points in this continuum are of potential interest. For instance, in one scenario, someone might want to get an idea of the types of things people are saying about a particular product through reading a sample of reviews covering the spectrum from highly positive, through balanced, to highly negative. (We call a review balanced if it is an opinionated text with an undecided or weak sentiment direction). In another scenario, a would-be customer might only be interested in reading balanced reviews, since they often present more reasoned arguments with fewer unsupported claims. Such a person might therefore want to avoid reviews such as Example (1) – written by a Chinese purchaser of a mobile phone (our English gloss).

(1)
软件不行，发送短信时有时对方接收不到；兼容性也不行，有的手机收到的短信是乱码！还有死机现象！拍照效果次！不是循环或自定义式闹铃，每次都要调，太麻烦了！后盖不够严密！原装配件中无座充！

The software is bad, some sent SMS are never received by the addressee; compatibility is also bad, on some mobile phones the received messages are in a scrambled encoding! And sometimes the phone 'dies'! Photos are horrible! It doesn't have a cyclic or pro-

grammable alarm-clock, you have to set it every time, how cumbersome! The back cover does not fit! The original software has many holes!

In a third scenario, someone might decide they would like only to read opinionated, weakly negative reviews such as Example (2), since these often contain good argumentation while still identifying the most salient bad aspects of a product.

(2)
这机子的反应速度超慢的哦，彩信必须要 30KB 以下才能收，也不支持 MP3 铃声，自带铃声也不好听，时不时的还会死机，本来买的时候挺喜欢的，样子挺独特，红色白色搭配的，挺有个性，也不贵，但是用着实在是总出状况，让人头疼

The response time of this mobile is very long, MMS should be less than 30kb only to be downloaded, also it doesn't support MP3 ring tones, (while) the built-in tunes are not good, and from time to time it 'dies', but when I was buying it I really liked it: very original, very nicely matching red and white colours, it has its individuality, also it's not expensive, but when used it always causes trouble, makes one's head ache

The review contains both positive and negative sentiment covering different aspects of the product, and the fact that it contains a balance of views means that it is likely to be useful for a would-be customer. Moving beyond review classification, more advanced tasks such as automatic summarization of reviews (e.g. Feiguina & LaPalme, 2007) might also benefit from techniques which could distinguish more shades of sentiment than just a binary positive / negative distinction.

A second dimension, orthogonal to positive / negative, is opinionated / unopinionated (or equivalently subjective / objective). When shopping for a product, one might be interested in the physical characteristics of the product or what features the product has, rather than opinions about how well these features work or about how well the product as a whole functions. Thus, if one is looking for a review that contains more factual information than opinion, one might be interested in reviews like Example (3).

(3)
总的感觉这台机器还不错，实用的有：开（关）机闹钟 5 个，800 条（500 个人）电话本，阴阳历显示，时间与日期快速转换，WAP 上网，日程表，记事本等。

(My) overall feeling about this mobile is not bad, it features: 5 alarm-clocks that switch the phone on (off), phone book for 800 items (500 people), lunar and solar calendars, fast switching between time and date modes, WAP networking, organizer, notebook and so on.

This review is mostly neutral (unopinionated), but contains information that could be useful to a would-be customer which might not be in a product specification document, e.g. fast switching between different operating modes. Similarly, would-be customers might be interested in retrieving completely unopinionated documents such as technical descriptions and user manuals. Again, as with sentiment classification, we argue that opinionated and unopinionated texts are not easily distinguishable separate sets, but form a continuum. In this continuum, intermediate points are of interest as well as the extremes.

A major obstacle for automatic classification of sentiment and objectivity is lack of training data, which limits the applicability of approaches based on supervised machine learning. With the rapid growth in textual data and the emergence of new domains of knowledge it is virtually impossible to maintain corpora of tagged data that cover all – or even most – areas of interest. The cost of manual tagging also adds to the problem. Reusing the same corpus for training classifiers for new domains is also not effective: several studies report decreased accuracy in cross-domain classification (Engström, 2004; Aue & Gamon, 2005) a similar problem has also been observed in classification of documents created over different time periods (Read, 2005).

In this paper we describe an unsupervised classification technique which is able to build its own sentiment vocabulary starting from a very small seed vocabulary, using iterative retraining to enlarge the vocabulary. In order to avoid problems of domain dependence, the vocabulary is built using text from the same source as the text which is to be classified. In this paper we work with Chinese, but using a very small seed vocabulary may mean that this approach would in principle need very little linguistic adjustment to be applied to a different

language. Written Chinese has some specific features, one of which is the absence of explicitly marked word boundaries, which makes word-based processing problematic. In keeping with our unsupervised, knowledge-poor approach, we do not use any preliminary word segmentation tools or higher level grammatical analysis.

The paper is structured as follows. Section 2 reviews related work in sentiment classification and more generally in unsupervised training of classifiers. Section 3 describes our datasets, and Section 4 the techniques we use for unsupervised classification and iterative retraining. Sections 5 and 6 describe a number of experiments into how well the approaches work, and Section 7 concludes.

2 Related Work

2.1 Sentiment Classification

Most previous work on the problem of categorizing opinionated texts has focused on the binary classification of positive and negative sentiment (Turney, 2002; Pang et al., 2002; Dave at al., 2003). However, Pang & Lee (2005) describe an approach closer to ours in which they determine an author's evaluation with respect to a multi-point scale, similar to the 'five-star' sentiment scale widely used on review sites. However, authors of reviews are inconsistent in assigning fine-grained ratings and quite often star systems are not consistent between critics. This makes their approach very author-dependent. The main differences are that Pang and Lee use discrete classes (although more than two), not a continuum as in our approach, and use supervised machine learning rather than unsupervised techniques. A similar approach was adopted by Hagedorn et al. (2007), applied to news stories: they defined five classes encoding sentiment intensity and trained their classifier on a manually tagged training corpus. They note that world knowledge is necessary for accurate classification in such open-ended domains.

There has also been previous work on determining whether a given text is factual or expresses opinion (Yu & Hatzivassiloglu, 2003; Pang & Lee, 2004); again this work uses a binary distinction, and supervised rather than unsupervised approaches.

Recent work on classification of *terms* with respect to opinion (Esuli & Sebastiani, 2006) uses a three-category system to characterize the opinion-related properties of word meanings, assigning numerical scores to Positive, Negative and Objective categories. The visualization of these scores somewhat resembles our graphs in Section 5, although we use two orthogonal scales rather than three categories; we are also concerned with classification of documents rather than terms.

2.2 Unsupervised Classification

Abney (2002) compares two major kinds of unsupervised approach to classification (co-training and the Yarowsky algorithm). As we do not use multiple classifiers our approach is quite far from co-training. But it is close to the paradigm described by Yarowsky (1995) and Turney (2002) as it also employs self-training based on a relatively small seed data set which is incrementally enlarged with unlabelled samples. But our approach does not use point-wise mutual information. Instead we use relative frequencies of newly found features in a training subcorpus produced by the previous iteration of the classifier. We also use the smallest possible seed vocabulary, containing just a single word; however there are no restrictions regarding the maximum number of items in the seed vocabulary.

3 Data

3.1 Seed Vocabulary

Our approach starts out with a seed vocabulary consisting of a single word, 好 (*good*). This word is tagged as a positive vocabulary item; initially there are no negative items. The choice of word was arbitrary, and other words with strongly positive or negative meaning would also be plausible seeds. Indeed, 好 might not be the best possible seed, as it is relatively ambiguous: in some contexts it means *to like* or acts as the adverbial *very,* and is often used as part of other words (although usually contributing a positive meaning). But since it is one of the most frequent units in the Chinese language, it is likely to occur in a relatively large number of reviews, which is important for the rapid growth of the vocabulary list.

3.2 Test Corpus

Our test corpus is derived from product reviews harvested from the website IT168[1]. All the reviews were tagged by their authors as either positive or negative overall. Most reviews consist of two or three distinct parts: positive opinions, negative opinions, and comments ('other') – although some

[1]http://product.it168.com

reviews have only one part. We removed duplicate reviews automatically using approximate matching, giving a corpus of 29531 reviews of which 23122 are positive (78%) and 6409 are negative (22%). The total number of different products in the corpus is 10631, the number of product categories is 255, and most of the reviewed products are either software products or consumer electronics. Unfortunately, it appears that some users misused the sentiment tagging facility on the website so quite a lot of reviews have incorrect tags. However, the parts of the reviews are much more reliably identified as being positive or negative so we used these as the items of the test corpus. In the experiments described in this paper we used 2317 reviews of mobile phones of which 1158 are negative and 1159 are positive. Thus random choice would have approximately 50% accuracy if all items were tagged either as negative or positive[2].

4 Method

4.1 Sentiment Classification

As discussed in Section 1, we do not carry out any word segmentation or grammatical processing of input documents. We use a very broad notion of words (or phrases) in the Chinese language. The basic units of processing are 'lexical items', each of which is a sequence of one or more Chinese characters excluding punctuation marks (which may actually form part of a word, a whole word or a sequence of words), and `zones', each of which is a sequence of characters delimited by punctuation marks.

Each zone is classified as either positive or negative based whether positive or negative vocabulary items predominate. In more detail, a simple maximum match algorithm is used to find all lexical items (character sequences) in the zone that are in the vocabulary list. As there are two parts of the vocabulary (positive and negative), we correspondingly calculate two scores using Equation (1)[3],

$$S_i = \frac{L_d}{L_{phrase}} S_d N_d \quad (1)$$

where L_d is the length in characters of a matching lexical item, L_{phrase} is the length of the current zone in characters, S_d is the current sentiment score of the matching lexical item (initially 1.0), and N_d is a negation check coefficient. The negation check is a regular expression which determines if the lexical item is preceded by a negation within its enclosing zone. If a negation is found then N_d is set to –1. The check looks for six frequently occurring negations: 不 (bu), 不会 (buhui), 没有 (meiyou), 摆脱 (baituo), 免去 (mianqu), and 避免 (bimian).

The sentiment score of a zone is the sum of sentiment scores of all the items found in it. In fact there are two competing sentiment scores for every zone: one positive (the sum of all scores of items found in the positive part of the vocabulary list) and one negative (the sum of the scores for the items in the negative part). The sentiment direction of a zone is determined from the maximum of the absolute values of the two competing scores for the zone.

This procedure is applied to all zones in a document, classifying each zone as positive, negative, or neither (in cases where there are no positive or negative vocabulary items in the zone). To determine the sentiment direction of the whole document, the classifier computes the difference between the number of positive and negative zones. If the result is greater than zero the document is classified as positive, and vice versa. If the result is zero the document is balanced or neutral for sentiment.

4.2 Iterative Retraining

The task of iterative retraining is to enlarge the initial seed vocabulary (consisting of a single word as discussed in Section 3.1) into a comprehensive vocabulary list of sentiment-bearing lexical items. In each iteration, the current version of the classifier is run on the product review corpus to classify each document, resulting in a training subcorpus of positive and a negative documents. The subcorpus is used to adjust the scores of existing positive and negative vocabulary items and to find new items to be included in the vocabulary.

Each lexical item that occurs at least twice in the corpus is a candidate for inclusion in the vocabulary list. After candidate items are found, the system calculates their relative frequencies in both the positive and negative parts of the current training subcorpus. The system also checks for negation while counting occurrences: if a lexical item is preceded by a negation, its count is reduced by one. This results in negative counts (and thus negative relative frequencies and scores) for those items that

[2]This corpus is publicly available at http://www.informatics.sussex.ac.uk/users/tz21/it168test.zip

[3]In the first iteration, when we have only one item in the vocabulary, negative zones are found by means of the negation check (so *not* + *good* = negative item).

are usually used with negation; for example, 质量太差了(*the quality is far too bad*) is in the **positive part** of the vocabulary with a score of –1.70. This means that the item was found in reviews classified by the system as positive but it was preceded by a negation. If during classification this item is found in a document it will reduce the positive score for that document (as it is in the positive part of the vocabulary), unless the item is preceded by a negation. In this situation the score will be reversed (multiplied by –1), and the positive score will be increased – see Equation (1) above.

For all candidate items we compare their relative frequencies in the positive and negative documents in the subcorpus using Equation (2).

$$difference = \frac{|F_p - F_n|}{(F_p + F_n)/2} \quad (2)$$

If *difference < 1,* then the frequencies are similar and the item does not have enough distinguishing power, so it is not included in the vocabulary. Otherwise the the sentiment score of the item is (re-) calculated – according to Equation (3) for positive items, and analogously for negative items.

$$\frac{F_p}{F_p + F_n} \quad (3)$$

Finally, the adjusted vocabulary list with the new scores is ready for the next iteration.

4.3 Objectivity Classification

Given a sentiment classification for each zone in a document, we compute sentiment density as the proportion of opinionated zones with respect to the total number of zones in the document. Sentiment density measures the proportion of opinionated text in a document, and thus the degree to which the document as a whole is opinionated.

It should be noted that neither sentiment score nor sentiment density are absolute values, but are relative and only valid for comparing one document with other. Thus, a sentiment density of 0.5 does not mean that the review is half opinionated, half not. It means that the review is less opinionated than a review with density 0.9.

5 Experiments

We ran the system on the product review corpus (Section 3.2) for 20 iterations. The results for binary sentiment classification are shown in Table 1. We see increasing F-measure up to iteration 18, after which both precision and recall start to decrease; we therefore use the version of the classifier as it stood after iteration 18[4]. These figures are only indicative of the classification accuracy of the system. Accuracy might be lower for unseen text, although since our approach is unsupervised we could in principle perform further retraining iterations on any sample of new text to tune the vocabulary list to it.

We also computed a (strong) baseline, using as the vocabulary list the NTU Sentiment Dictionary (Ku et al., 2006)[5] which is intended to contain only sentiment-related words and phrases. We assigned each positive and negative vocabulary item a score of 1 or –1 respectively. This setup achieved 87.77 precision and 77.09 recall on the product review corpus.

In Section 1 we argued that sentiment and objectivity should both be considered as continuums, not

Iteration	Precision	Recall	F-measure
1	77.62	28.43	41.62
2	76.15	73.81	74.96
3	81.15	80.07	80.61
4	83.54	82.79	83.16
5	84.66	83.78	84.22
6	85.51	84.77	85.14
7	86.59	85.76	86.17
8	86.78	86.11	86.44
9	87.15	86.32	86.74
10	87.01	86.37	86.69
11	86.9	86.15	86.53
12	87.05	86.41	86.73
13	86.87	86.19	86.53
14	87.35	86.67	87.01
15	87.13	86.45	86.79
16	87.14	86.5	86.82
17	86.8	86.24	86.52
18	87.57	86.89	87.22
19	87.23	86.67	86.95
20	87.18	86.54	86.86

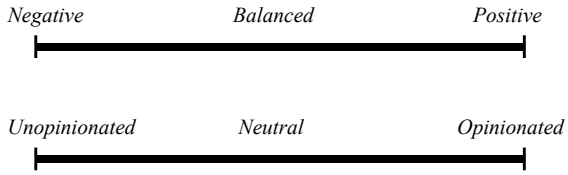
Table 1. Results for binary sentiment classification during iterative retraining.

[4]The size of the sentiment vocabulary after iteration 18 was 22530 (13462 positive and 9068 negative).

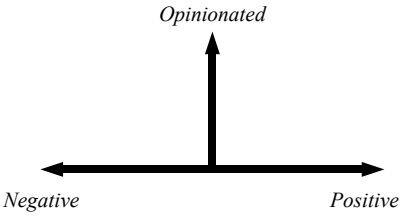
[5]Ku et al. automatically generated the dictionary by enlarging an initial manually created seed vocabulary by consulting two thesauri, including tong2yi4ci2ci2lin2 and the Academia Sinica Bilingual Ontological WordNet 3.

binary distinctions. Section 4.1 describes how our approach compares the number of positive and negative zones for a document and treats the difference as a measure of the 'positivity' or 'negativity' of a review. The document in Example (2), with 12 zones, is assigned a score of –1 (the least negative score possible): the review contains some positive sentiment but the overall sentiment direction of the review is negative. In contrast, Example (1) is identified as a highly negative review, as would be expected, with a score of –8, from 11 zones.

Similarly, with regard to objectivity, the sentiment density of the text in Example (3) is 0.53, which reflects its more factual character compared to Example (1), which has a score of 0.91. We can represent sentiment and objectivity on the following scales:



The scales are orthogonal, so we can combine them into a single coordinate system:



We would expect most product reviews to be placed towards the top of the the coordinate system (i.e. opinionated), and stretch from left to right.

Figure 1 plots the results of sentiment and objectivity classification of the test corpus in this two dimensional coordinate system, where *X* represents sentiment (with scores scaled with respect to the number of zones so that –100 is the most negative possible and +100 the most positive), and *Y* represents sentiment density (0 being unopinionated and 1 being highly opinionated).

Most of the reviews are located in the upper part of the coordinate system, indicating that they have been classified as opinionated, with either positive or negative sentiment direction. Looking at the overall shape of the plot, more opinionated documents tend to have more explicit sentiment direction, while less opinionated texts stay closer to the balanced / neutral region (around *X = 0*).

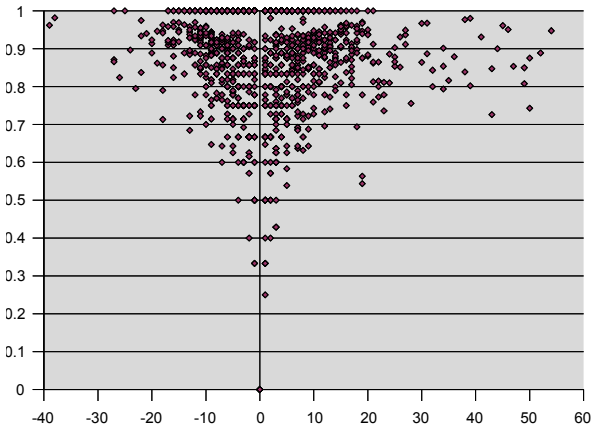


Figure 1. Reviews classified according to sentiment (*X* axis) and degree of opinionation (*Y* axis).

6 Discussion

As can be seen in Figure 1, the classifier managed to map the reviews onto the coordinate system. However, there are very few points in the neutral region, that is, on the same *X = 0* line as balanced but with low sentiment density. By inspection, we know that there are neutral reviews in our data set. We therefore conducted a further experiment to investigate what the problem might be. We took Wikipedia[6] articles written in Chinese on mobile telephony and related issues, as well as several articles about the technology, the market and the history of mobile telecommunications, and split them into small parts (about a paragraph long, to make their size close to the size of the reviews) resulting in a corpus of 115 documents, which we assume to be mostly unopinionated. We processed these documents with the trained classifier and found that they were mapped almost exactly where balanced documents should be (see Figure 2).

Most of these documents have weak sentiment direction (*X = –5* to *+10*), but are classified as relatively opinionated (*Y > 0.5*). The former is to be expected, whereas the latter is not. When investigating the possible reasons for this behavior we noticed that the classifier found not only feature descriptions (like 手感很好 *nice touch*) or expressions which describe attitude (喜欢 *(one) like(s)*), but also product features (for example, 彩信 *MMS* or 电视 *TV*) to be opinionated. This is because the presence of some advanced features such as MMS in mobile phones is often regarded as a positive by

[6]www.wikipedia.org

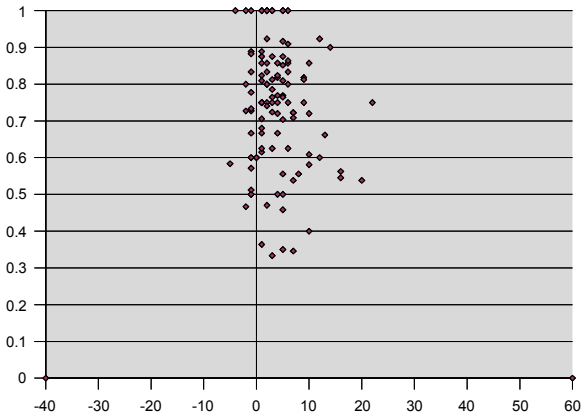


Figure 2. Classification of a sample of articles from Wikipedia.

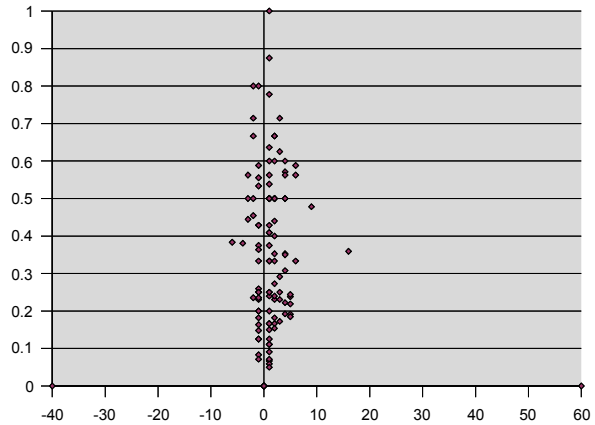


Figure 3. Classification of a sample of articles from Wikipedia, using the NTU Sentiment Dictionary as the vocabulary list.

authors of reviews. In addition, the classifier found words that were used in reviews to describe situations connected with a product and its features: for example, 服务 (*service*) was often used in descriptions of quite unpleasant situations when a user had to turn to a manufacturer's post-sales service for repair or replacement of a malfunctioning phone, and 用户 (*user*) was often used to describe what one can do with some advanced features. Thus the classifier was able to capture some product-specific as well as market-specific sentiment markers, however, it was not able to distinguish the context these generally objective words were used in. This resulted in relatively high sentiment density of neutral texts which contained these words but used in other types of context.

To verify this hypothesis we applied the same processing to our corpus derived from Wikipedia articles, but using as the vocabulary list the NTU Sentiment Dictionary. The results (Figure 3) show that most of the neutral texts are now mapped to the lower part of the opinionation scale ($Y < 0.5$), as expected. Therefore, to successfully distinguish between balanced reviews and neutral documents a classifier should be able to detect when product features are used as sentiment markers and when they are not.

7 Conclusions and Future Work

We have described an approach to classification of documents with respect to sentiment polarity and objectivity, representing both as a continuum, and mapping classified documents onto a coordinate system that also represents the difference between balanced and neutral text. We have presented a novel, unsupervised, iterative retraining procedure for deriving the classifier, starting from the most minimal size seed vocabulary, in conjunction with a simple negation check. We have verified that the approach produces reasonable results. The approach is extremely minimal in terms of language processing technology, giving it good possibilities for porting to different genres, domains and languages.

We also found that the accuracy of the method depends a lot on the seed word chosen. If the word has a relatively low frequency or does not have a definite sentiment-related meaning, the results may be very poor. For example, an antonymous word to 好 (*good*) in Chinese is 坏 (*bad*), but the latter is not a frequent word: the Chinese prefer to say 不好 (*not good*). When this word was used as the seed word, accuracy was little more than 15%. Although the first iteration produced high precision (82%), the size of the extracted subcorpus was only 24 items, resulting in the system being unable to produce a good classifier for the following iterations. Every new iteration produced an even poorer result as each new extracted corpus was of lower accuracy.

On the other hand, it seems that a seed list consisting of several low-frequency one-character words can compensate each other and produce better results by capturing a larger part of the corpus (thus increasing recall). Nevertheless a single word may also produce results even better than those for multiword seed lists. For example, the two-character word 方便 (*comfortable*) as seed reached 91%

accuracy with 90% recall. We can conclude that our method relies on the quality of the seed word. We therefore need to investigate ways of choosing 'lucky' seeds and avoiding 'unlucky' ones.

Future work should also focus on improving classification accuracy: adding a little language-specific knowledge to be able to detect some word boundaries should help; we also plan to experiment with more sophisticated methods of sentiment score calculation. In addition, the notion of 'zone' needs refining and language-specific adjustments (for example, a 'reversed comma' should not be considered to be a zone boundary marker, since this punctuation mark is generally used for the enumeration of related objects).

More experiments are also necessary to determine how the approach works across domains, and further investigation into methods for distinguishing between balanced and neutral text.

Finally, we need to produce a new corpus that would enable us to evaluate the performance of a pre-trained version of the classifier that did not have any prior access to the documents it was classifying: we need the reviews to be tagged not in a binary way as they are now, but in a way that reflects the two continuums we use (sentiment and objectivity).

Acknowledgements

The first author is supported by the Ford Foundation International Fellowships Program.

References

Abney, Steven (2002) Bootstrapping. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, Philadelphia, PA. 360–367.

Aue, Anthony & Michael Gamon (2005) Customizing sentiment classifiers to new domains: a case study. In *Proceedings of RANLP-2005*.

Dave, Kushal, Steve Lawrence & David M. Pennock (2003) Mining the peanut gallery: opinion extraction and semantic classification of product reviews. In *Proceedings of the Twelfth International World Wide Web Conference*. 519–528.

Engström, Charlotte (2004) *Topic dependence in sentiment classification*. Unpublished MPhil dissertation, University of Cambridge.

Esuli, Andrea & Fabrizio Sebastiani (2006) SENTIWORDNET: a publicly available lexical resource for opinion mining. In *Proceedings of LREC-06, the 5th Conference on Language Resources and Evaluation*, Genoa, Italy.

Hagedorn, Bennett, Massimiliano Ciaramita & Jordi Atserias (2007) World knowledge in broad-coverage information filtering. In *Proceedings of the 30th ACM SIGIR Conference on Research and Development in Information Retrieval*. 801–802.

Ku, Lun-Wei, Yu-Ting Liang & Hsin-Hsi Chen (2006) Opinion extraction, summarization and tracking in news and blog corpora. In *Proceedings of the AAAI-2006 Spring Symposium on Computational Approaches to Analyzing Weblogs*, AAAI Technical Report.

Feiguina, Olga & Guy Lapalme (2007) Query-based summarization of customer reviews. In *Proceedings of the 20th Canadian Conference on Artificial Intelligence*, Montreal, Canada. 452–463.

Pang, Bo, Lillian Lee & Shivakumar Vaithyanathan (2002) Thumbs up? Sentiment classification using machine learning techniques. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Philadelphia, PA. 79–86.

Pang, Bo & Lillian Lee (2004) A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, Barcelona, Spain. 271–278.

Pang, Bo & Lillian Lee (2005) Seeing stars: exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, Ann Arbor, MI. 115–124.

Read, Jonathon (2005) Using emoticons to reduce dependency in machine learning techniques for sentiment classification. In *Proceedings of the Student Research Workshop at ACL-05*, Ann Arbor, MI.

Turney, Peter D. (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*, Philadelphia, PA. 417–424.

Yarowsky, David (1995) Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, Cambridge, MA. 189–196.

Yu, Hong & Vasileios Hatzivassiloglou (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*, Sapporo, Japan. 129–136.