Unsupervised sentiment classification of English movie reviews using automatic selection of positive and negative sentiment items

John Rothfels Julie Tibshirani

June 2, 2010

Abstract

We consider the problem of classifying documents not by topic, but by overall sentiment. Previous approaches to sentiment classification have favored domain-specific, supervised machine learning (Naive Bayes, maximum entropy classification, and support vector machines). Inherent in these methodologies is the need for annotated training data. Building on previous work, we examine an unsupervised system of iteratively extracting positive and negative sentiment items which can be used to classify documents. Our method is completely unsupervised and only requires linguistic insight into the semantic orientation of sentiment.

1 Introduction

Humans love categorization. Robert Sapolsky, a professor of biology at Stanford University who studies human behavior, explains that categorizing or "bucketing" (no intentional homage to hashing, alas) information enables the human brain to process information in a more meaningful and natural way. As an example, humans may not be able to tell the difference between two beams of light at a wavelengths of 510nm and 511nm, but it is easy and natural to call each "green". The green "bucket" thus encompasses a range of wavelengths which through a single word we may call the "same", more or less.

Today, very large amounts of information are publicly available online. As part of an effort to give organization to this sprawling mess of data, researchers have studied the problem of text categorization, i.e. assigning a label to text on the basis of topical subject. More recent studies have focused on the subject of sentiment classification, i.e. assigning a label to text on the basis of the overall sentiment of the author (*positive* or *negative*, for example). The application of such research is immediate. Often what people care about when reading a review, for example, is not what the review says exactly but whether it is holistically positive or negative. This type of classification can be used not only in the consumer sector, but also in business sector for product recommendation or inflammatory message filtering. In general, we see that sentiment classification allows us to extract something quantitative out of a vast amounts of qualitative information.

Pang, Lee, and Vaithyanathan [7] concluded that sentiment classification is inherently more difficult than text categorization, finding as one example that a common phenomenon in classifying the sentiment of films was a "thwarted expectations" narrative, where the author sets up a deliberate contrast to earlier discussion: "This film should be brilliant. It sounds like a great plot, the actors are first grade, and the supporting cast is good as well, and Stallone is attempting to deliver a good performance. However, it cant hold up". More generally, we acknowledge that the subjective nature of sentiment makes classification particularly difficult (even human annotators can have difficulty detecting the sentiment of certain text).

On the spectrum of sentiment, we make one necessary but grossly simplifying assumption: that sentiment is binary (i.e *positive* or *negative*). The assumption, while deliberate, results in a serious problem, namely that the difference between sentiment "buckets", to use Sapolsky's term, is magnified. Two documents differing only slightly in sentiment (one ever so slightly positive and one ever so slightly negative) will appear markedly dissimilar by our classification scheme, in much the same way that an artificial wavelength boundary separates blue from green while two beams of light on either side of this boundary look identical. Thus at some level, sentiment classifica-

tion is an all-too-artificial construct. These issues acknowledged, we press on.

2 Related Work

2.1 Supervised Methods

Supervised approaches to sentiment classification have attracted quite a bit of recent attention. Pang, Lee, and Vaithyanathan [7] compared multiple supervised machine learning algorithms (Naive Bayes, maximum entropy classifiers, support vector machines) for the task of sentiment classification of movie reviews. They experimented with a wide variety of features and obtained accuracy as high as 82.9% when classifying movie reviews. Nevertheless, they were not able to achieve accuracies on sentiment classification comparable to those reported for standard topic-based categorization. Many typical misclassifications were easily recognizable by humans. Experimentation with feature selection has manifested in other research: Dave, Lawrence, and Pennock [2] experimented with the use of linguistic, statistical, and n-gram features and measures for feature selection and weighting. Pang and Lee [5] used a graph-based technique to identify and analyze only subjective parts of texts. Yu and Hatzivassiloglou [9] use semantically oriented words for identification of polarity at the sentence level.

In all supervised approaches, reasonably high accuracy can be obtained subject only to the requirement that test data be similar to training data. To move a supervised sentiment classifier to another domain would require collecting annotated data in the new domain and retraining the classifier. This dependency on annotated training data is one major shortcoming of all supervised methods.

2.2 Unsupervised Methods

Unsupervised approaches to sentiment classification can solve the problem of domain dependency and reduce the need for annotated training data. Turney [8] uses two arbitrary seed words (*poor* and *excellent*) to calculate the semantic orientation of phrases, where the orientation of a phrase is defined as the difference of its association with each of the seed words (as measured by pointwise mutual information). The sentiment of a document is calculated as the average semantic orientation of all such phrases. This approach was able to achieve 66% accuracy for the movie review domain at the document level.

Zagibalov and Carroll [10] describe a method of automatic seed word selection for unsupervised sentiment classification of product reviews in Chinese. The method only requires information about commonly occurring negations and adverbials in order to iteratively find sentiment bearing items. The results obtained are close to those of supervised classifiers and sometimes better, up to an F_1 score of 92%. We discuss their strategy in detail in the next sections.

3 Our Approach

Similar to Zagibalov and Carroll [10], our main goal is to overcome the problem of domain dependency in sentiment classification. Unsupervised approaches seem promising in this regard since they do not require annotated training data, just access to sufficient raw text in each domain. We wanted to look at a problem focused enough to be able to delve deeply into error analysis, rather than spend time presenting a broad survey of unsupervised machine learning techniques and their results. Our reasoning is twofold: (1) that most academic papers suggest that unsupervised methods will *not* perform as well as supervised methods no matter how many we try, and (2) that survey papers are trite.

Our choice is to continue the work of Zagibalov and Carroll [10]. Their suggestions for future work are to explore issues of positivity of language use (they find that most sentiment seeds extracted through their unsupervised method are positive), as well as to see if their system works equally well for languages other than Chinese and in domains which contain evaluative language (for example, movie reviews). We proceed by adapting their approach to English and evaluating its performance on the movie review domain.

3.1 The Movie Review Domain

For our experiments, we chose to work with movie reviews. This domain is experimentally convenient because the data has already been compiled and made publicly available by Pang, Lee, and Vaithyanathan [7]. We work with 2000 movie reviews, 1000 for each sentiment bucket (*positive* and *negative*). Since there has already been substantial work conducted in the domain (mostly supervised machine learning), we had abundant resources with which to compare our findings. To our satisfaction, we found no studies other

than Turney's [8] that have attempted unsupervised sentiment classification in the movie domain. We note that Turney found movie reviews to be the most difficult of several domains for sentiment classification, but we stress that our experiments are not specific to movie reviews and should be easily applicable to other domains with sufficiently large corpuses. Our hope is that by focusing on movie reviews (notoriously difficult due to prolific sarcasm and subnarrative structure) we will see a lower bound on performance.

4 Method and Experiments

4.1 Overview

We adapted the approach of Zagibalov and Carroll [10] for unsupervised sentiment classification. The intuition behind their approach is that "positive sentiment seeds" can be extracted from text on the basis of occurring frequently after negation, but more frequently without negation. A "positive sentiment seed" is defined as a sequence of characters (their research uses Chinese text, an inconsistency which we consider below), i.e. a lexical item, which at a reductive level expresses positive sentiment. A sentiment of a document can then be approximated by the sentiment seeds contained within it. With a set of initial positive sentiment seeds we can iteratively classify the documents and use the information from this classification to expand the list of positive sentiment seeds, as well as extract negative sentiment seeds. Presumably, classification will improve as we extract more sentiment seeds.

The initial set of positive sentiment seeds is extracted based on the following intuition:

1. Attitude is often expressed through the negation of vocabulary items with the opposite meaning; for example, it is more common to say *not good* than *bad*. Zagibalov and Carroll's system uses this observation to find negative lexical items while nevertheless starting only from a positive seed (*good*).

2. The polarity of a candidate seed needs to be determined. We assume that the lexical item *good* can be used as a gold standard for positive lexical items and compare the pattern of contexts a candidate seed occurs in to the pattern exhibited by the gold standard.

The study finds that *good* appears with negation, but more frequently without it. With this insight, we can extract positive sentiment seeds on the basis of following the same pattern. The study cites an overall "positivity of language" to describe this pattern. As an additional constraint proposed by the paper, positive sentiment seeds should only be extracted if they start with an adverb, to avoid extracting seeds without content (i.e. *not a boy* occurring less frequently than *a boy* should not lead us to suspect that *a boy* indicates positive sentiment).

4.2 Initial Strategy

To form an initial set of positive seeds, we perform two passes over the data. We first extract all adverbial phrases of a set length (e.g four words) that follow a negation ("not", "isn't", etc.), then prune this list to contain only those phrases that occur in our corpus more frequently without negation than with.

Proceeding with Zagibalov and Carroll's protocol, we use this set of positive sentiment seeds to classify each "zone" of a document. A "zone" in Chinese text is defined as a sequence of characters between punctuation. The score of a zone is calculated from the score of each positive sentiment seed contained within it by the following equation:

$$\sum \frac{S_i \cdot N_i}{L_{zone}}$$

Here, S_i is the current weight of the lexical item (initially 1.0), L_{zone} is the length of the zone, and N_i is the negation coefficient for the i-th lexical item of the zone, -1 if it is preceded by a negation, 1 otherwise. A zone is classified as *positive* if its score is positive, *negative* otherwise, and the sentiment of an entire document is *positive* if there are more positive zones than negative zones.

With an initial classification for each document, we add to our list of sentiment seeds every lexical item which appears significantly more often in one type of document than the other. Let F_p and F_n be the frequency of a lexical item in positive and negative documents. Then an item appears significantly more often in one classification over the other if

$$\frac{2 \cdot |F_p - F_n|}{F_p + F_n} > 1$$

The score of the newly included lexical item is $F_p - F_n$. Note that this step has the potential to introduce negative sentiment seeds if a seed occurs significantly

more often in negative documents than positive. In this case, $F_p - F_n$ is negative and the sentiment seed will contribute a negative weight in classifying zones (unless preceded by a negation). With an expanded set of sentiment seeds, we lather, rinse, and repeat until there are no changes to the classification of any single document between iterations.

4.2.1 Analysis and Results

To our dismay, our initial results were very unimpressive compared to a baseline classifier (a classifier which produces binary output uniformly at random has an accuracy of 50% in expectation). Our classifier achieved only minimal improvement from the baseline, reporting accuracy of 50.3%. There were clearly deep-rooted issues with our model, which we hypothesized was as a combination of implementation specifics and language differences between Chinese and English. Examining our model, we first found that splitting documents on punctuation to produce "zones" is possibly a poor choice in English, while it seems to make more sense in Chinese where characters have no internal punctuation. Many of the zones we produced were nonsensical or did not carry any content ("by now", "however", "the next day", "as previously stated"). When we stopped splitting the data into zones, however, we found that many of the seeds extracted included punctuation, which we felt made them less semantically meaningful and less likely to appear multiple times in the corpus (thus increasing the sparsity of our seed set).

In addition, scoring documents different zones (rather than overall score) produced a common problem: very frequently, there would be an equal number of positive zones and negative zones. The initial study uses this strategy to prevent one extremely positive zone (i.e. very positively weighted) from exerting too much influence over the classification of the entire document. However, we found that we could achieve better results by summing over the scores given to each zone of a document, rather than simply counting the number of positive and negative zones, when determining its polarity. As our own security measure, we include a zone length normalization factor into the scoring for each zone. With this improvement in place, we managed to boost accuracy to 51.3%.

Examining the length of the seeds we extracted, we found that 4-grams seem to capture a lot of content because they are long enough to include the adjective and then the noun it modifies after the negation. However, it often stops right at a preposition and so confusingly includes a proposition while not capturing the prepositional attachment. Also, 4-grams are notoriously sparse, and many observed seeds included words specific to the movie which was being reviewed, such as actor's names and characters. Thus we experiment with trigrams and bigrams, but found that we had lost much of the content of phrases, so we begrudgingly stuck with 4-grams.

Because of the requirement that we need to have seen a seed at least three times before including it in our initial set of positive seeds (once preceded by a negation and two times without for a positive difference), we miss many seeds which have content and instead seem to focus instead on common language constructions ("to take advantage of", "the beginning of the", "going to have to", "to do with the") as opposed to contentful seeds ("to waste your hard-earned", "even worth a 99-cent", "really blow me away").

Still, at this point we found that the 4-grams our model selected as seeds were quite poor. There seems to be a deep issue with the linguistic intuition used to pick these seeds. Looking at the words after a negation and an adverbial is awkward because often times the phrase contains many structural words or else is the beginning of a complicated sentence with a verb and subject. Many of the most semantically useful phrases actually had a negative tone, as demonstrated in the samples above. However, we were stuck with 4-grams since smaller models produced non-contentful seed units, but this in turn (for reasons described above) had disastrous consequences in selecting useful positive sentiment seeds. It seemed clear to us at this point that adapting Zagibalov and Carroll's methodology to English sentiment classification suffers from severe language differences, most notably that Chinese uses implicit structure while succinctly capturing sentiment while English is much more verbose and variable.

4.3 A Second Attempt

Our initial analysis seems to validate the intuition that sentiment is best captured through small lexical units, often adjectives or small adjective or adverbial phrases (e.g. "multi-dimensional characters", "remotely sympathetic"), see Turney [8]. Consequently, we revised our system to look for small, semantically meaningful seeds. Our intuition is that these positive seeds ("good", "wonderful", "surprising") will occur in much the same pattern as larger lexical items (such

as our earlier 4-grams), either standalone in the positive case or directly following negation or a negated adverbial in the negative case.

4.3.1 Analysis and Results

A small shift in approach amounted to a small change in our algorithm, which was able to produce, among many others, the following initial seeds: "interesting", "hilarious", "prominent", "inventive". This list seemed promising. Our iterative approach to extracting additional sentiment seeds and reweighting preexisting ones, however, seemed to be ineffectual. Running our classifier with and without iteration still produced accuracy right around the baseline (51.5% and 50.8%, respectively). Looking at the changes between scores from iterations beyond the first, we found that very few seeds are weighted significantly different than they were in previous iterations, which suggests that documents are being classified almost identically between iterations, even as we continue to extract additional sentiment seeds. This is, of course, contrary to our preliminary expectations.

We suspected that the issue may be in the scoring of documents, not so much in the selection of sentiment seeds (some of these seemed to be, for the most part, pretty good). To test our theory, as an alternative scoring mechanism we tried k-means clustering on the set of movie reviews, where each review is translated into a vector with one element per sentiment seed. For each document, the value for a given element is the number of times the corresponding sentiment seed is found in the document times the score given to the sentiment seed, accounting as always for negation coefficient. Running the algorithm, however, produced nearly identical output (accuracy 50.9%), or just over baseline.

4.4 Final Method

Our near complete failure at adapting Zagibalov and Carroll's classification scheme led us to question the legitimacy of the linguistic assumptions that form the foundation of their approach. In particular it seems as though "positivity of language", the idea underlying our choice of final positive seeds as being sufficient to extract other seeds, does not apply to English. Our own revised approach to choosing the initial seeds, which restricts our set of adverbs and extracts adjective unigrams, does produce seeds that are strong indicators of sentiment. However, the majority of these seeds are not clearly positive ("unbearable" and "ineffective" appeared alongside "insightful" and "dramatic"). We also lose many strongly positive seeds when pruning the list of potential seeds; our method eliminates such words as "amazing", "entertaining", and "sincere", among others.

With these doubts in mind, we decided to adopt a more linguistically-neutral approach to choosing seed words. For this we turned to the unsupervised classification scheme proposed by Turney [8], which has been proven (at least moderately) effective in classifying English movie reviews. The system uses a metric called the "semantic orientation" of a phrase, which is defined as the difference of the phrase's association with positive and negative seed words (as measured by pointwise mutual information):

$$SO(p) = PMI(p, excellent) - PMI(p, poor)$$

The PMI terms of the equation were, in Turney's study, calculated through queries to a database (each hit counts as a co-occurrence of the two words) where the phrases considered were bigrams consisting of an adjective and a noun, thus making it more likely for them to have semantic content. A document is classified as positive if the sum over the semantic orientations of its terms is greater than zero, and negative otherwise.

We adapted this idea of semantic orientation to choose the initial set of seeds, while opting to keep Zagibalov and Carroll's iterative classification algorithm. The algorithm works as follows:

1. Hand pick two sets of reference seeds, one positive and one negative. We used for our positive set, somewhat arbitrarily, the words "good", "excellent", "amazing", "incredible", and "great"; and for our negative set, "bad", "poor", and "terrible".

2. Pass through the data calculating the semantic orientation between each word in the corpus (required to be a single adjective) and the sets P and N (*positive* and *negative*). Since we now compare against sets instead of single words, we use a revised formula for PMI where a co-occurrence of a word with the positive set, for example, occurs whenever the word appears in the same document as any single word in the set (weighted by the number of times a word from the positive set appears in the document).

3. Keep only those seeds that have strong semantic weight (we chose $|SO| > 1$) to form the final list of seeds.

4. Perform the iterative classification approach described earlier, using each term's semantic orientation as its initial sentiment score.

4.4.1 Analysis and Results

The reasons for our revision are twofold: (1) that we avoid relying on repeated access to a large internet database (a known bias in Turney's study), and (2) by calculating semantic orientations from the corpus itself, our seeds are more likely to reflect the semantic patterns of the particular domain. With the adapted algorithm, our results improved dramatically: classifying with just our initial set of positive seeds we were able to achieve a 65.5% accuracy on the set of labelled movie reviews.

Surprisingly, though, our accuracy showed no improvement upon iterative reclassification of the documents. During the first iteration, the number of seeds in the set jumps from 813 to over 2500, suggesting that many of these words are not semantically meaningful, and upon inspecting the list of added seeds we do in fact observe that these words are largely neutral or rare adjectives. Our hypothesis is that adding and rescoring seeds provides little additional information, and simply serves to solidify the initial classification. We were pleased to see, at least, that iterative reclassification through extraction of additional sentiment items did not hurt our overall accuracy.

5 Conclusions

This paper sought to adapt Zagibalov and Carroll's unsupervised sentiment classification system from Chinese product reviews to the domain of English movie reviews. Our results were consistently unimpressive, leading us to question the linguistic assumptions underlying their choice of initial positive seeds. Although it is common for sentiment-bearing items to occur after negations, we found it was not safe in our domain to assume "positivity of language", i.e. that positive seeds occur distinctly more without negation than with, and that in general there is more positivity than negativity as, perhaps, a function of socialization. It seems that this assumption does not apply equally well to English as Chinese, and moreover that it favors particular domains of application (movie reviews are notoriously snarky, and so negative adjectives may dominate). Of course, domain specificity is exactly what we were trying to avoid with our classification scheme. In addition, English phrases contain more syntax and verbosity than their Chinese counterparts, and so extracting the phrase after a negation produces a seed with a low ratio of length to content.

These realizations led us to adopt a new approach to choosing initial sentiment seeds, one that focused on small semantic units and allowed us to better tailor the choice to the domain. We blended Turney's approach, which uses PMI-IR to determine the polarity of small adjective phrases, with Zagibalov and Carroll's iterative reclassification algorithm. Our results were modest but significantly better than before: we were able to achieve an accuracy of 65.5% on our set of movie reviews, as opposed to a near-baseline accuracy of 50.3% with the original approach. The iterative reclassification algorithm, however, failed to improve the classifications, and future work must be done to fix its faults.

6 Future Work

In his study, Turney [8] notes that for movie reviews, "the whole is not necessarily the sum of the parts". His suggestion is that often times, such as with the "thwarted expectation" narrative, we will have documents of one polarity which contain phrases distinctly of another polarity. Our scoring scheme, however, is very much based on a "sum of the parts" intuition. Our primary suggestion for future work is to explore alternate scoring schemes given our method of extracting sentiment items. One idea is to incorporate discourse structure into our classification. An example of this is subjectivity analysis, i.e. determining when somebody is talking about the plot of a movie rather than their subjective opinion. Parts which are *not* subjective can interfere with a "sum of the parts" scoring approach, since the only "parts" we really want to consider are those which are subjective. A pertinent example would be a section of a movie review describing a horribly gruesome scene, but which nonetheless belongs to a positive review. Another idea is to incorporate sentiment flow into our classification scheme, which aims to model the global sentiment of a document as a trajectory of local sentiments. This can help identify "thwarted narrative" type reviews by helping the classifier to understand more globally what the sentiment of the document is.

As one final suggestion, we reconsider the benefits of supervised methods. Our own goal to avoid domain specificity and the need for annotated training data is important; however, supervised approaches across the board tend to perform much better. Consequently, we suggest a bootstrapping approach which uses our method to identify sentiment seeds and run an initial classification of documents and then proceeds to use supervised machine learning for a more robust classifier.

References

[1] A. Andreevskaia and S. Bergler. Mining WordNet for fuzzy sentiment: Sentiment tag extraction from WordNet glosses. In *Proceedings of EACL*, volume 6, 2006.

[2] K. Dave, S. Lawrence, and D.M. Pennock. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In *Proceedings of the 12th international conference on World Wide Web*, page 528. ACM, 2003.

[3] S. Ghosh and M. Koch. Unsupervised Sentiment Classification Across Domains.

[4] V. Hatzivassiloglou and K.R. McKeown. Predicting the semantic orientation of adjectives. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, page 181. Association for Computational Linguistics, 1997.

[5] B. Pang and L. Lee. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the ACL*, volume 2004, 2004.

[6] B. Pang and L. Lee. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135, 2008.

[7] B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 79–86. Association for Computational Linguistics, 2002.

[8] P.D. Turney et al. 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*, pages 417–424, 2002.

[9] H. Yu and V. Hatzivassiloglou. Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In *Proceedings of EMNLP*, volume 3, pages 129–136, 2003.

[10] T. Zagibalov and J. Carroll. Automatic seed word selection for unsupervised sentiment classification of Chinese text. In *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*, pages 1073–1080. Association for Computational Linguistics, 2008.