

# Improving Sentimental Classifications Using Contextual Sentences

Nathan Sakunkoo  
Department of Computer Science  
Stanford University  
Stanford, CA 94305 USA  
[sakunkoo@stanford.edu](mailto:sakunkoo@stanford.edu)

## **Abstract**

This paper presented a new methodology, which helps improve the accuracy of sentimental polarity classification. Unlike most prior works which focused on lexical features at the word level, the methodology presented here attempts to include more contextual information by focusing on the sentence level. This paper proposes the following process: (1) train a classifier using word-level lexical features, (2) use the classifier to label each sentence in the document, (3) train another classifier based on the labeled sentences and classify accordingly.

## **1. Introduction**

In recently years, there has been rapid growth in online reviews and discussion groups, in which an important characteristic of the entries is their sentiments or judgmental opinions on some subjects. Many researchers have attempted to classify an opinion as positive or negative, an analysis called sentimental classification. At first glance, sentimental classification is very similar to topical classification; however, it is still rarely the case that sentimental classification could achieve the same performance as topical classification. In some respects, the sentimental classification poses significant distinctive challenges to current researchers. While topics are usually identifiable by using only keywords, sentiments are rarely straightforward; they are normally expressed in a more subtle manner. For example, the sentence “*Why would anyone watch this movie?*” contains no obvious negative words, but it implies negativity. Hence, sentimental

classification seems to require more contextual analysis than topical classification.

Another specific problem pointed out by Pang, Lee, and Vaithyanathan (Pang et al., 2002) is the “thwarted exceptions,” where the author sets up a deliberate contrast to his/her earlier discussion; therefore, although he expresses many positive features in the text, his opinion should, in fact, be classified as a negative opinion. This example, again, demonstrates that sentimental classification seems to require a deeper understanding and treatment than topical classification.

Despite its importance and strong potential applications (Pang et al., 2002), there were still limited research findings in this area. This paper aimed to solve such problems by trying to include more contextual information at the “sentence” level.

Traditional approaches had typically focused on selecting indicative word-level lexical features (i.e. “good”). In contrast, this paper attempts to extend the bag-of-words model by incorporating more contextual information, the approach that is natural in human’s reading comprehension. The broad idea is that, in order to help alleviate problems such as Pang’s thwarted expectations, a classifier was further trained sentence-by-sentence. Thus, this paper proposes the following process: (1) train a classifier using word-level lexical features, (2) use the classifier to label each sentence in the document, (3) train another classifier based on the labeled sentences and then classify the whole document (opinion) accordingly. This

sentence-level approach, in addition to the word-level one, should prevent our polarity classifier from being too word literal and being misled by the number of words. Thus, it should alleviate the thwarted expectations problem and result in higher accuracy. To the author's knowledge, no previous work has integrated this analytical approach thus far.

My results show that this “*Bag of Sentences*” approach helped increase the accuracy of the polarity classification task approximately 2-3% from the baseline models, a statistically significant improvement. Furthermore, my combination of this approach with Pang & Lee's *Subjectivity Summarization* further boosted the accuracy up at an additional 5-6% from the baseline models.

## 2. Previous Works

This research project primarily builds on the works of Pang & Lee's EMNLP 2002 and Pang & Lee's ACL 2004.

Pang & Lee (2002) examined whether it suffices to treat sentiment classification simply as a special case of topic-based categorization (with two topics representing different polarities). They conducted sentimental classification using three standard classification techniques (Naïve Bayes, Maximum Entropy, Support Vector Machine). Finally, they examined several factors that make sentimental classification more challenging such as the thwarted expectation problem mentioned in the prior section. Turney (2002) also made a similar note that in sentimental context, “the whole is not necessarily the sum of the parts”

In 2004, Pang and Lee further proposed a novel method that applied text-categorization techniques to only the subjective portions of the documents. This Subjectivity Summarization method helped increase accuracy about 3-4%.

## 3. Data

Movie reviews are typically subjective and can be classified as either positive or negative opinions. They are thus suitable for our classification evaluation. My experiments use Pang & Lee's movie review datasets from <http://www.cs.cornell.edu/People/pabo/movie-review-data/>.

### *Sentiment Polarity Datasets (v 2.0)*

The data source was gathered from the Internet Movie Database IMDB archive of [rec.arts.movies.reviews](http://rec.arts.movies.reviews). There are 1000 positive and 1000 negative full text movie reviews. The ground truths (actual polarity ratings in this case) were assigned based on Pang & Lee's heuristic scripts extracting the first review score from each review such as Thumbs Up/ Thumbs Down and Star Ratings.

### *Subjectivity Datasets (v 1.0)*

The data source contains 5000 *objective* sentences and 5000 *subjective* sentences. Objective sentences were extracted from Internet Movie Database's plot summaries while the subjective sentences are from Rotten Tomatoes' review snippets.

## 4. Approach

This paper aims to improve polarity classification of the default bag of words model. To help comprehend the overall tone of the text, this paper suggests classifying the sentiment using sentences as inputs, the “Bag of Sentences” model.

The experiments are summarized in 4.1-4.5.

### 4.1 Bag of Sentences (Standard Model)

#### *Procedure 4a*

Step 1: Train Word-Level Classifier using n-grams as inputs.

Step 2: Use Word-Level Classifier to classify each sentence as positive or negative.

Step 3: Train Sentence-Level Classifier using the label of each sentence (positive/negative)

Step 4: Test the accuracy of both Word-Level and Sentence-Level Classifier

#### 4.2 Bag of Sentence (Weighted Ratio Heuristic)

*Procedure 4b* (using Average Ratio of positive/negative sentences instead of NB classifiers at the Sentence Level)

Step 1: Same as Procedure 4a.

Step 2: Same as Procedure 4a.

Step 3: Instead of training the Classifier by using sentences as input, this procedure simply calculates the average number of positive/negative reviews.

Step 4: Use the average number of positive/negative sentences as a threshold in classifying reviews. Then apply the classifier on the test set.

#### 4.3 Subjectivity Summary Extraction

Pang & Lee's ACL 2004 showed that using Subjectivity Summary helped increase accuracy about 3-4%. The steps are as follow.

*Procedure 4c*

Step 1: Train a Classifier on Subjectivity Corpus

Step 2: Store the Subjectivity Model for further use.

*Procedure 4d*

Step 1: Train classifier at word-level using n-grams as inputs. (Polarity dataset)

Step 2: Extract top 5 subjective sentences (or up to 25 if the likelihood to be subjective > 50%) by using the Subjectivity Classifier from Procedure 4c.

Step 3. Test the extracted subjective sentences against the classifier in Step 1.

In addition, a more sophisticated form (4.4-4.5) that combines 4.1 or 4.2 with 4.3 may further improve our polarity classification task.

#### 4.4 Subjective Sentences Extraction + Bag of Sentence Standard

This approach combines 4.1 and 4.3 together. The procedure is same as Procedure 4a except that during Step 2,3,4, train and test only subjective sentences.

#### 4.5 Subjective Sentences Extraction + Bag of Sentence (Ratio Heuristic)

This approach combines 4.2 and 4.3 together.

### **5. Classifier Model**

Our aim in this work is to examine whether looking at the sentence level would help improve sentiment polarity classification. The author started the experiment by using the Naïve Bayes (NB) classifier. NB is chosen because of its simplicity and effectiveness. Another classifier called Language Model is also used. The latter was chosen because it is interesting and is readily available in the software LingPipe NLP library that is used in this experiment.

#### 5.1 Naïve Bayes Classifier

$$P(c | d) = \frac{P(c)P(d | c)}{P(d)}$$

Using Bayes' rule above, a Naïve Bayes (NB) classifier will assign a given document  $d$  the class  $c^* = \operatorname{argmax}_c P(c|d)$ .

$$P_{\text{NB}}(c | d) := \frac{P(c) (\prod_{i=1}^m P(f_i | c)^{n_i(d)})}{P(d)}$$

Since  $P(d)$  is constant,  $\operatorname{argmax}_c P(c|d)$  is the same as  $P(c,d)$ .

The NB Classifier implemented in LingPipe uses Character Sequences as inputs rather than Word Sequences.

## 5.2 Language Model Classifier

The LMClassifier from LingPipe performs joint probability-based classification based on the language models for each category and a multivariate distribution over categories.

Similar to NB Classifier, LM Classifier will assign a given document  $d$  the class  $c^* = \operatorname{argmax}_c P(c|d) = \operatorname{argmax}_c P(c,d)$  where  $P(c,d) = P(d|c) * P(c)$

## 6. Experiments and Results

### 6.1 Experiment Setup

The author used the documents from the movie-review data described in Section 3. The dataset contains 1000 positive reviews and 1000 negative reviews. 400 reviews (200 positive and 200 negatives) were held out as test data. The other 1,600 were treated as training data.

The dataset was extracted from the original HTML documents. Explicit rating indicators (such as 10/10) were removed.

No stemming or stop words were used. The unigram implemented here is pure unigram. This means that “not good” was treated as two different unigrams although, ideally, in unigram cases, negation terms should be combined with the term that followed, as implemented in Pang et. al 2002.

### 6.2 Results

Below, the author first presents initial results and analysis of the initial results, and then proceeds to report the main results.

**Initial Results** (Approach 4.1: Bag of Sentences Standard Model with both NB and LM from 5.1 and 5.2)

The results of the NB Classifier and the LM Classifier are presented in Figures 6a and 6b, respectively.

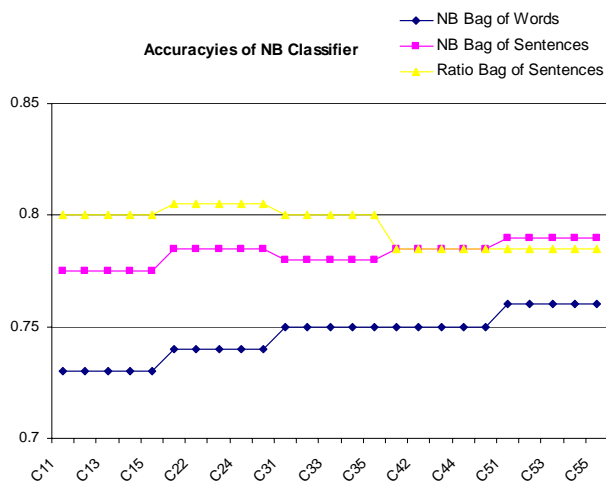


Figure 6a  
Figure 6a:  $C_{xy}$  means using  $X$ -gram Language Model for Word-Level and using  $Y$ -gram during Sentence-Level

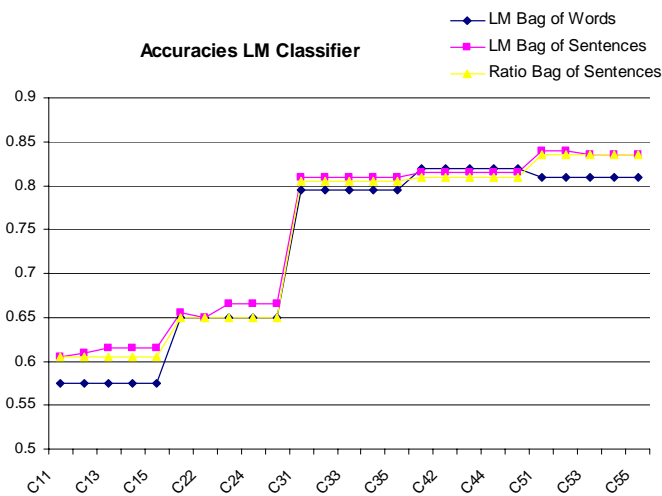


Figure 6b

As shown in Figure 6a and Figure 6b, in most cases, the Classifier using the Bag of Sentences models outperformed the Classifier using Bag of Words model by about 2 percent.

One rationale behind the improvement is that, from a look at the opinions, we can see that each sentence typically has its own tone (assuming subjective sentence/ clause); thus two sentences normally represent the reviewer’s sentiment more affirmatively and

accurately than one. Therefore, when the bag of sentences analysis was conducted, the sentences are analyzed and the big picture was achieved. However, the bag of words model only analyzed at the word levels, without regards of the number of sentences. For example, as shown in Figure 6c below, the reviewer affirmed his opinion with the sentence “I think not” twice (6 words). We can see that these two sentences both attested to his opinion that he doesn’t think so. The bag of sentences model would interpret the two sentences as 2 negatives out of 2 sentences, a 100% negativity. The bag of words model, on the other hand, would yield 2 negatives out of 6 words, a mere 33% negativity. This is an important reason why the bag of sentences model yields greater accuracy.

```
Classify as pos,neg vs neg (BoW,BoS vs ground truth)
nPos:4  nNeg:9
****Words****
pos don't let this movie fool you into believing the romantic n
neg no one will truly understand the heart and soul of this man
neg any moves to ? glamorise' his life , which hollywood has an
pos this movie about his life , although well written , puts to
pos oh well , let's fantasise onwards an assume that he was a b
neg it is easier for me to believe that he had a wet dream and
neg but i guess my version probably wouldn't draw a crowd or ma
neg so is there any justification in romanticising the man shak
neg i think not .
pos as for the oscars were they deserved by this movie ?
neg i think not .
neg in many aspects ? private ryan' and ? life is beautiful' we
neg another sore point is the fact that gwyneth won the best fe:
```

Figure 6c

Alternative Approach (Pre-Approach 4.2: Bag of Sentence (Simple Ratio Heuristic))

From the error analysis of the experiments, I observed that in many of the cases where the Bag of Sentences (BoS) model performed better than the Bag of Words (BoW) model, the classification could also be easily determined by comparing the numbers of positive sentences with negative sentences. Therefore, the author tried using a simpler heuristic of classifying the reviews by classifying each review as positive if the number of positive sentences is greater than the number of negative sentences. The results were quite poor and are not reported here to avoid distracting the overall picture. An example of the poorer result is shown in

Figure 6d below, where there are equal numbers of positive and negative sentences, but the opinion should be classified as positive. Thus, refinement is needed.

```
Classify as neg,pos,pos VS pos (BoW,BoS,Ratio VS ground truth)
nPos:9  nNeg:9
****Words****
pos originally titled 'don't lose your head' , this parody of the
neg two english fops , the 'powdered , be-wigged , be-ribboned' si
pos due to a series of machinations and disguises , they are large
neg ffig becomes known as 'the black fingernail' because he leave
neg after the fingernail rescues a prominent royalist the duc de p
neg ( in fact , darcy and ffig are their coachmen ! )
pos once at calais , the fingernail meets jacqueline ( dany robin
pos he tells her his identity and gives her his locket .
neg when camembert realises that the fingernail is nearby , he sea
neg jacqueline is imprisoned in the bastille and camembert , his l
neg they pretend to be of noble stock , calling themselves the duc
neg desiree finds out that ffig is the fingernail by wearing the
pos ffig attempts to stall camembert so that he can return to the
pos a more complex story than most carry ons , this film enjoys go
neg sid james is excellent as the english fop and black fingernail
pos other acting honours go to joan sims who is perfect as desiree
pos although it suffers from a disasterously over-long sword fight
pos definitely one of the best of the series and a joy to watch .
```

Figure 6d

Refinement: (Approach 4.2: Bag of Sentence (Weighted Ratio Heuristic))

After a closer examination of the data and generated output, the author observed that positive sentences should carry more weight than do negative sentences. It is interesting to note that most reviews in the dataset contain more negative sentences than positive sentences. This might be because the reviewers generally tend to be more elaborative on the negative sides than on positive sides. I also observed that that there existed a sizeable number of the reviews in which the reviewers described their negative perceptions of the movies extensively or highlighted many poor aspects of the movies, but strongly loved some particular aspects of the movies, and ended up rating the movies positively.

Hence, Approach 4.2 was introduced. Instead of having a fixed threshold of classifying a document as positive when #positive sentence>=#negative sentence (#positive/#negative >= 1.0), Approach 4.2 uses total #positive sentence/ total #negative sentence as the threshold. Roughly, negative documents have the ratio of 0.5 and positive documents have the ratio of 0.9 on average.

Logically, the threshold used for classifying movie reviews in this project fluctuates around 0.7. This bias towards negative comments coincides with our prior observation that negative sentences typically outnumbered positive sentences whether or not the reviewers like the movies.

The performance of Approach 4.2 is marginally better than 4.1 with some fluctuation in NB cases. The fact that this simple heuristic at the evaluation step of the classification surprisingly performed on par and even slightly better than the training bag of sentences implies that there is room for improvement in training bag of sentences, an interesting direction for future research. Please see future direction in Section 7.

### Subjectivity Summarization

Next, Approach 4.3 (Subjectivity Summarization) was tested. The accuracy levels are in line with Pang and Lee’s studies (2004). In this study, this approach increased the performance 3.5% for NB classifiers and 2% for LM classifiers. This contrasts with Approach 4.1 and 4.2, where LM achieved slightly higher performance than NB. The rationale behind this finding is still unclear at present. It seems possible that the LM classifier built separate language models for positive and negative documents and thus helped labeling the sentiment for each sentence more accurately. The accuracy at the word level obviously affected the labeling of a sentence more than it affected the labeling of the whole document.

Compared to the Subjectivity Summarization model, our Bag of Sentence model performed better on LM classifiers and on par (sometimes slightly worse) on NB classifiers. Again, the difference could be due to the nature of the two classifiers as described earlier.

Finally, the integrative approaches of 4.4 and 4.5 were conducted. They combined the BoS

approach in 4.1 and 4.2 with the subjectivity summarization in 4.3, respectively. The results are shown in Figure 6e and 6f. From the graphs, we can see that the integrative approach resulted in an even higher level of accuracy, approaching 87% in the case of 5-gram LM model of BoS with Ratio Heuristic and Subjectivity Summary.

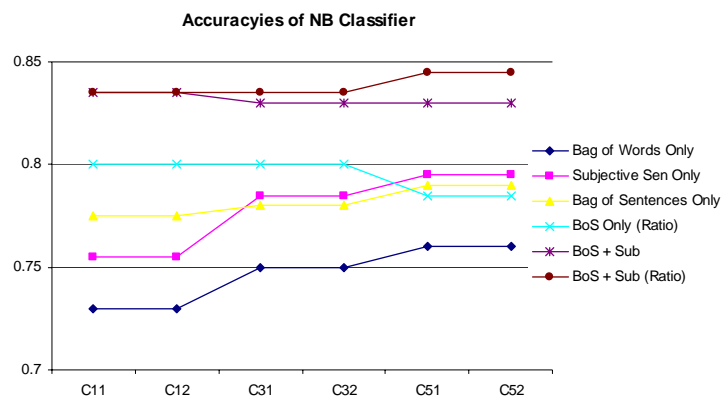


Figure 6e Cxy means using X-gram Language Model during Word-Level and using Y-gram during Sentence-Level Analysis

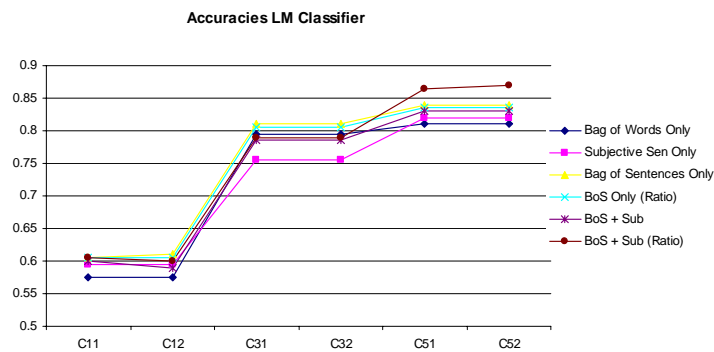


Figure 6f

```

Classify as pos VS neg (actual)
nPos:13 nNeg:7
****Words***
pos make no mistake , deep rising has anaconda beat all to
pos here , the laughs are provided almost entirely by kevin
pos it's hard to work up much enthusiasm for this sort of :
pos as it happens , deep rising is noteworthy primarily fo
neg it's also gloomy , uninspired and not nearly enough fun
pos i delighted in the sneaky-smart entertainment of ron ur
pos deep rising is missing that one unmistakable cue that i
neg sure enough , all the modern monster movie ingredients
neg i don't ask much of my monster movies , but i do ask tl
pos writer/director stephen sommers seems most concerned w
pos you don't have to show me a fantastically impressive ,
neg just show me the massive beast burping , and i'll figu
pos there may not be a critic alive who harbors as much af
pos he shrieks in horror at his freakish appearance and pa
pos deep rising , a big-undersea-serpent yarn , doesn't qu
pos me , characters ; an isolated location , here a dereli
pos case it point , comparing deep rising to its recent co
neg he looks at another character . . . and
neg winks .
neg and up go the lights .
-----

```

Figure 6g

### Discussion for Improvement

Our model showed significant improvement for sentimental classification. However, based on this study, even our best result for the sentimental classification at 87% still falls behind that of the topic-based classification (approximately 90%) by a modest gap. This is in line with prior research that showed the subtle difficulty of sentimental classification. A detailed error analysis of the results showed that our “Bag of Sentences” model still does not specifically solve the “Thwarted Expectation” problem as originally expected by the author (see Figure 6g). The result suggests that BoS does include more contextual information than BoW, and hence more effective; however, it alone is not enough to solve the thwarted expectation problem. Perhaps adding weights to different zones of the document and also taking argument indicators into account might help reduce the thwarted expectation problem.

```

184 Classify as pos VS neg (actual)
185 nPos:4 nNeg:2
186 ****Words***
187 pos while the animation is certainly colorful to look at , osmosis jones
188 pos while far less offensive than the farrelly's last effort " me , myse
189 neg the story cries out for puny puns , but we only get occasional sprin
190 pos one must wonder how the climatic flatlining of a child's father will
191 neg " osmosis jones " will probably be ok for the kids , but the farrel
192 pos rest assured , the whole enchilada is wrapped up with a fart joke .
193 -----

```

Figure 6h

I also found that, while the global context poses difficulty for sentimental polarity classification especially the “thwarted

expectation” problem, local contextual problems such as “word sense ambiguity” during sentence labeling also challenge the accuracy of sentimental classification. For instance, the last line in Figure 6h was labeled incorrectly as a positive sentence. This is primarily because of the words “rest assured” and “joke”. In many other reviews, these two phrases are good indicators of positive reviews (especially in the comedy genre). Unfortunately, this is not the case in the above movie’s genre–terror. On a side note, the first few lines of Figure 6h also give an example of the occasional failure to deal with some aspects of the “thwarted expectation” problem.

Furthermore, approaches 4.1-4.3 were tested at different sizes of training data to further ensure robustness. Figure 6i shows that accuracy increases monotonically (given the data points) as training sizes increase.

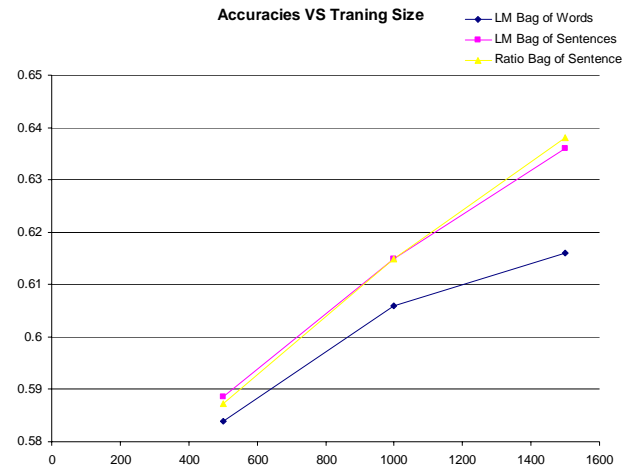


Figure 6i

## 7. Conclusion and Future Work

This paper presented a new methodology, which effectively helps to improve the accuracy of sentimental polarity classification by approximately 3-5%. Unlike most prior works that focused on lexical features at the word level, the methodology presented here tries to include more contextual information

by also analyzing the sentence level. In short, this paper proposes the following process: (1) train a classifier using word-level lexical features, (2) use the classifier to label each sentence in the document, (3) train another classifier based on the labeled sentences (or possibly combine with Subjectivity Summary as in 4.4-4.5), and then classify accordingly.

My experimental results showed that the Bag of Sentences systems performed better than the baseline Bag of Words model with statistical significance, in almost all experimented cases (Figure 6a, 6b, 6e, 6f). My approach in 4.1-4.2 by itself also performed on par with Pang and Lee's Subjectivity Summarization method on Naïve Bayes classifiers, and even slightly better on LM classifiers.

We also improved the performance by combining two complementary methods—the Bag of Sentences and the Subjectivity Summarization—, which led to improvement in sentimental polarity classification accuracy.

### Future Directions

One original purpose of my “Bag of Sentences” method was to solve the “thwarted expectation” problem. Although it did not directly solve the problem entirely, it did help reduce such kind of errors and capture more contextual information for the classification process. In future research, the author plans to continue to build on this study in order to further solve the “thwarted expectation” problem by using zones as mentioned earlier in Section 6. More generally, one could also try to extract sentences that contain summaries of the sentiments (using document structure, the zones, or keywords such as “overall” or “in sum”).

Another area that the author would like to work on is to increase the external validity of the approach by performing more comprehensive tests. Currently, the dataset contains 2000 documents, the number that is

still somewhat vulnerable to idiosyncrasies. In addition, the author hopes to implement the classifiers from the ground up, so that more customization can be achieved, and so that the results could be directly compared with Pang & Lee's paper. It would also be interesting to try the “Bag of Sentences” method with MaxEnt and SVM classifiers, especially during the evaluation phase in which the bags of sentences are classified as positive or negative. As pointed out earlier in Section 6, the higher performance of our weighted Ratio Heuristic suggested that this phase has potential for improvement.

### **Acknowledgement**

The author would like to acknowledge alias-i's LingPipe. The implementation of the experiment in this paper relies heavily on the LingPipe library and its tutorial document and code.

The author would like to thank Patty Min Sakunkoo for discussion and grammatical and logical corrections of this paper.

Finally, the author would like to thank Professor Christopher Manning and the TAs for the insightful classes and discussions.

### **Reference**

Bo Pang and Lillian Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In Proc. 42nd ACL.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? sentiment classification using machine learning techniques. In Proceedings of EMNLP.

Peter D. Turney. 2002. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In Proceedings 40th Annual Meeting of the ACL (ACL'02),

Note: Some figures were chopped off due to the picture capturing problem.