# Enhancing Lexicon-Based Review Classification by Merging and Revising Sentiment Dictionaries

**Heeryon Cho**
Yonsei Institute of Convergence Technology, Yonsei University
Incheon, Republic of Korea

**Jong-Seok Lee** **Songkuk Kim**
School of Integrated Technology
Yonsei University
Incheon, Republic of Korea

{heeryon, jong-seok.lee, songkuk}@yonsei.ac.kr

## Abstract

This paper presents a method of improving lexicon-based review classification by merging multiple sentiment dictionaries, and selectively removing and switching the contents of merged dictionaries. First, we compare the positive/negative book review classification performance of eight individual sentiment dictionaries. Then, we select the seven dictionaries with greater than 50% accuracy and combine their results using (1) averaging, (2) weighted-averaging, and (3) majority voting. We show that the combined dictionaries perform only slightly better than the best single dictionary (65.8%) achieving (1) 67.8%, (2) 67.7%, and (3) 68.3% respectively. To improve this, we combine seven dictionaries at a deeper level by merging the dictionary entry words and averaging the sentiment scores. Moreover, we leverage the skewed distribution of positive/negative threshold setting data to update the merged dictionary by selectively removing the dictionary entries that do not contribute to classification while switching the polarity of selected sentiment scores that hurts the classification performance. We show that the revised dictionary achieves 80.9% accuracy and outperforms both the individual dictionaries and the shallow dictionary combinations in the book review classification task.

## 1 Introduction

With the increase in opinion mining and sentiment analysis-related researches, various lexical resources that define sentiment scores/categories have been constructed and made available. Examples include SentiSense (de Albornoz *et al.*, 2012), SentiWordNet (Baccianella *et al.*, 2010), Micro-WNOp, and WordNet-Affect (Strapparava and Valitutti, 2004), which are based on a large English lexical database WordNet (Fellbaum, 1998), and AFINN (Nielsen, 2011), Opinion Lexicon (Hu and Liu, 2004), Subjectivity Lexicon (Riloff and Wiebe, 2003) and General Inquirer (Stone and Hunt, 1963), which are manually or semi-automatically constructed. These resources differ in their formats and sizes, but all can be utilized in the lexicon-based opinion mining and sentiment analysis.

The increase in the number of sentiment resources naturally gives rise to two questions: (1) How are the performances of these resources different? (2) Can we construct a better sentiment resource by combining and/or revising multiple resources? We answer these questions by comparing the book review classification performance of single and combined sentiment resources, and present a simple 'merge, remove, and switch' approach that revises the entries of the sentiment resource to improve its classification performance.

In the next section, we describe the experimental setup for evaluating the classification performance of sentiment resources. We then compare the positive/negative classification performance of eight widely known individual sentiment resources in Section 3. Since individual sentiment resources are originally constructed in different formats, we standardize their formats. These standardized resources will be called *sentiment dictionaries* or simply *dictionaries* throughout this paper. In Section 4 we compare the classification performances of combined dictionaries which integrate multiple individual dictionaries' results using averaging, weighted-averaging, and majority voting. Then, we introduce a method of revising sentiment dictionaries at a deeper level by merging, removing, and switching the dictionary contents. Implications for utilizing multiple dictionaries are discussed. Related works are introduced in Section 5, and conclusion is given in Section 6.



*International Joint Conference on Natural Language Processing*, pages 463–470,
Nagoya, Japan, 14-18 October 2013.

## 2 Experimental Setup

90,000 Amazon book reviews were collected to construct a positive/negative review dataset for sentiment dictionary evaluation.

### 2.1 Dataset

5-star and 4-star book reviews were merged and labeled as positive reviews, and 1-star and 2-star reviews were merged and labeled as negative reviews. 3-star reviews were excluded. 10,000 reviews (positive reviews: 9,007 / negative reviews: 993) were randomly selected as positive/negative threshold setting data (see 2.3). The remaining 80,000 reviews (positive reviews: 71,993 / negative reviews: 8,007) were set aside as test data.

### 2.2 Review Sentiment Score Calculation

Eight sentiment resources (see Table 1) were standardized to generate eight sentiment dictionaries ($D_j$, $j$=1,…,$d$). (The standardization of sentiment resources is discussed in Section 3.1.)

Each book review was tokenized, lemmatized, and part-of-speech tagged using the Stanford CoreNLP suite (Toutanova *et al.*, 2003). Once the list of words in the review was obtained, the sentiment score ($D_j(w_i)$) of each word was looked up in the sentiment dictionary ($D_j$). The scores of all the review words listed in the dictionary ($w_i$, $i$=1,…,$n$) were averaged to yield the *Review Sentiment Score* (*RSS*).

$$RSS(D_j) = \frac{1}{n}\sum_{i=1}^{n} D_j(w_i)$$

Because the Stanford Part-of-Speech Tagger outputs detailed parts of speech whereas the standardized sentiment dictionaries either do not define or define only four parts of speech (e.g., noun, adjective, verb, and adverb), the Tagger's parts of speech were mapped to four parts of speech as shown in Table 3.

| Resource | Entry Size | Sentiment Category & Score Range | Note |
|---|---|---|---|
| AFINN[1] | 2,477 words | No categories. Each word has integer score ranging between -5 (very negative) and 5 (very positive). | Based on Affective Norms for English Words (ANEW). |
| General Inquirer[2] | 11,788 words | Positiv/Negativ/Pstv/Ngtv/ Pleasur/Pain/EMOT/etc. categories. No numerical scores. | Based on Harvard IV-4 and Lasswell dictionaries, etc. |
| Micro-WNOp[3] | 1,105 synsets/ 1,960 words | Positive/negative/objective categories each with 0~1 score. | Based on WordNet 2.0. |
| Opinion Lexicon[4] | 6,786 words | Positive/negative categories. No numerical scores. | Misspelled words are deliberately included. |
| SentiSense[5] | 2,190 synsets/ 4,404 words | Joy/sadness/love/hate/despair/ hope/etc. 14 emotion categories. No numerical scores. | Based on WordNet 2.1. |
| SentiWordNet[6] | 117,659 synsets/ 155,287 words | Positive/negative/objective categories each with 0~1 score. | SentiWordNet ver. 3.0. Based on WordNet 3.0. |
| Subjectivity Lexicon[7] | 8,221 words | Positive/negative/both/neutral categories. No numerical scores. | Subjectivity (weak/ strong) is also defined. |
| WordNet-Affect[8] | 2,872 synsets/ 4,552 words | Synsets are first categorized into emotion/mood/trait/behavior/ etc., and these categories are further categorized into positive/negative/ambiguous /neutral. No numerical scores. | Based on WordNet 1.6. |

Table 1. The contents of eight sentiment resources.

[1] http://www2.imm.dtu.dk/pubdb/views/publication_details.php?id=6010
[2] http://www.wjh.harvard.edu/~inquirer/spreadsheet_guide.htm
[3] http://www-3.unipv.it/wnop/
[4] http://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html#lexicon
[5] http://nlp.uned.es/~jcalbornoz/resources.html
[6] http://sentiwordnet.isti.cnr.it/
[7] http://mpqa.cs.pitt.edu/lexicons/subj_lexicon/
[8] http://wndomains.fbk.eu/wnaffect.html

| | AFN[1] | GI[2] | MWO[3] | OL[4] | SS[5] | SWN[6] | SL[7] | WNA[8] |
|---|---|---|---|---|---|---|---|---|
| **AFN[1]** | 2,477<br>**2,454**<br>*1,723* | | | | | | | |
| **GI[2]** | **917**<br>*913* | 11,788<br>**3,906**<br>*3,853* | | | | | | |
| **MWO[3]** | **196**<br>*190* | **551**<br>*551* | 1,960<br>**1,515**<br>*1,334* | | | | | |
| **OL[4]** | **1,315**<br>*1,148* | **2,504**<br>*2,485* | **470**<br>*465* | 6,786<br>**6,560**<br>*5,393* | | | | |
| **SS[5]** | **771**<br>*742* | **1,238**<br>*1,237* | **397**<br>*375* | **1,533**<br>*1,476* | 4,404<br>**3,729**<br>*3,225* | | | |
| **SWN[6]** | **1,781**<br>*1,615* | **3,870**<br>*3,836* | **1,504**<br>*1,330* | **5,386**<br>*5,080* | **3,715**<br>*3,217* | 155,287<br>**77,761**<br>*33,923* | | |
| **SL[7]** | **1,246**<br>*1,182* | **3,047**<br>*3,021* | **586**<br>*582* | **5,296**<br>*4,771* | **1,738**<br>*1,685* | **6,130**<br>*5,860* | 8,221<br>**6,731**<br>*6,059* | |
| **WNA[8]** | **312**<br>*292* | **391**<br>*391* | **122**<br>*118* | **835**<br>*550* | **938**<br>*786* | **1,024**<br>*857* | **639**<br>*609* | 4,552<br>**1,035**<br>*864* |

Table 2. Number of shared single word entries disregarding the parts of speech between two dictionaries (**bold numbers**). Numbers in *italics* are the actual dictionary entries that match the book review words. The underlined numbers in the diagonal cells are the actual entry word size of each dictionary.

[1]AFINN, [2]General Inquirer, [3]Micro-WNOp, [4]Opinion Lexicon, [5]SentiSense, [6]SentiWordNet, [7]Subjectivity Lexicon, [8]WordNet-Affect.

### 2.3 Threshold Setting & Judgment

Each dictionary's threshold for judging the positivity and negativity (i.e., the review label) of the book reviews was set using the threshold setting data; the threshold with the greatest accuracy was selected. A review was judged as positive if the *RSS* was greater than or equal to the threshold, and as negative, otherwise.

$$ReviewLabel = \begin{cases} \text{positive} & \text{if } RSS \geq threshold \\ \text{negative} & \text{otherwise} \end{cases}$$

### 2.4 Performance Measure

Since the book review dataset was an imbalanced dataset containing more positive reviews than negative reviews (i.e., 9:1 ratio), balanced accuracy ($Acc_{BAL}$) was used to measure the overall performance.

$$Acc_{BAL} = 0.5 \times Recall_{POS} + 0.5 \times Recall_{NEG}$$

$$Recall_{POS} = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}}$$

$$Recall_{NEG} = \frac{\text{true negatives}}{\text{true negatives} + \text{false positives}}$$

$Recall_{POS}$ and $Recall_{NEG}$ each measure the positive and negative review accuracy.

| Stanford POS Tagger | Senti.Dict. |
|---|---|
| NN, NNS, NNP, NNPS | Noun |
| JJ, JJR, JJS | Adjective |
| VB, VBD, VBG, VBN, VBP, VBZ | Verb |
| RB, RBR, RBS | Adverb |

Table 3. Part-of-speech mapping from Stanford POS Tagger to the sentiment dictionary.

## 3 Individual Dictionary Comparison

The bold numbers in Table 2 indicate the number of shared *single words* between two sentiment dictionaries; note that the part-of-speech was disregarded when extracting the shared words. The underlined numbers in the diagonal cells are the actual dictionary entry word sizes. The italicized numbers are the dictionary entry words that actually match the book review words.

The eight sentiment dictionaries in Table 2 all include the following thirty-one words: "approval", "cheer", "cheerful", "contempt", "cynical", "disdain", "earnest", "excitement", "fantastic", "glee", "gloomy", "good", "guilt", "horrible", "marvel", "offend", "proud", "reject", "scorn", "sick", "sincerity", "sore", "sorrow", "sorry", "triumph", "trouble", "ugly", "upset", "vile", "warm", and "worry".

| Dictionary | $Recall_{POS}$ | $Recall_{NEG}$ | $Acc_{BAL}$ | Threshold |
|---|---|---|---|---|
| AFINN | 59.6% | 67.9% | 63.8% | 0.140 |
| General Inquirer | 62.2% | 69.3% | **65.8%** | 0.175 |
| Micro-WNOp | 40.7% | 70.0% | 55.4% | 0.120 |
| Opinion Lexicon | **67.0%** | 63.4% | 65.2% | 0.025 |
| SentiSense | 61.2% | 64.3% | 62.8% | 0.225 |
| SentiWordNet | **67.0%** | 64.3% | 65.7% | 0.005 |
| Subjectivity Lexicon | 58.7% | **70.6%** | 64.7% | 0.170 |
| WordNet-Affect | 59.8% | 38.2% | 49.0% | 0.005 |

Table 4. Positive ($Recall_{POS}$), negative ($Recall_{NEG}$), and balanced accuracy ($Acc_{BAL}$) of eight sentiment dictionaries on 80,000 book reviews.

### 3.1 Standardization

Because some sentiment resources define sentiment *categories* instead of sentiment *scores*, they were converted to sentiment scores: For example, *positive*, *negative*, and *neutral/ambiguous* categories were each converted to 1.0, -1.0, and 0.0.

In some cases, emotion categories such as *joy*, *sadness*, *love*, etc. were first mapped to *positive*, *negative*, or *ambiguous* categories and then converted to sentiment scores. The standardization process for each dictionary is explained below.

**AFINN:** AFINN contains sentiment scores ranging between $-5 \leq score_{AFINN} \leq 5$. These scores were normalized from [-5..5] to [-1..1].

Normalizing [A..B] to [C..D] employed the following equation:

$$X' = \frac{D-C}{B-A} \cdot X + \frac{C \times B - A \times D}{B-A}$$

The below equation was used for AFINN:

$$X' = 0.2X$$

**General Inquirer (GI):** Each entry word in the GI contains one or more GI categories, and we selected the following sentiment-related categories and calculated the sentiment scores by averaging the assigned category values: *Positiv*, *Pstv*, *PosAff*, *Pleasur*, *Virtue*, *Complet*, and *Yes* categories were each assigned a 1.0 score while *Negativ*, *Ngtv*, *NegAff*, *Pain*, *Vice*, *Fail*, *No*, and *Negate* categories were assigned a -1.0 score.

**Micro-WNOp:** For each entry word, the positive/negative paired sentiment scores were given by multiple human judges. These paired scores were added and averaged to obtain a single sentiment score. Note that for all WordNet-based sentiment resources, the different senses of a word (e.g., happy#1, happy#2, etc.) were aggregated and their sentiment scores were averaged.

**Opinion Lexicon:** Words in the positive word list were given a 1.0 score while words in the negative word list were given a -1.0 score. Three ambiguous words that were included in both the positive and negative lists were given a 0.0 score.

**SentiSense:** Emotional categories assigned to the synsets were converted to sentiment scores: *Joy*, *love*, *hope*, *calmness*, and *like* categories were given a 1.0 score; *fear*, *anger*, *disgust*, *surprise*, and *anticipation* categories were given a -1.0 score; and *ambiguous*, *surprise*, and *anticipation* categories were given a 0.0 score.

**Subjectivity Lexicon:** *Positive*, *negative* and *neutral* categories were converted to 1.0, -1.0, and 0.0 sentiment scores respectively. Entry words with 'anypos' (i.e., any parts-of-speech) were unfolded to have four parts-of-speech.

**WordNet-Affect:** Synsets having affective hierarchical categories such as *positive-emotion*, *negative-emotion*, *ambiguous-emotion*, and *neutral-emotion* were converted to 1.0, -1.0, 0.0, and 0.0 sentiment scores respectively.

Note that only the single word dictionary entries were actually looked up in the book review classification experiments; phrases or compound words (e.g., those including blank spaces, hyphens or underscores) were not matched.

### 3.2 Evaluation

The *RSS*s were calculated using the eight standardized sentiment dictionaries for each review, and the threshold for judging the review label was set differently for each dictionary using the 10,000 book review threshold setting data.

Table 4 compares the classification performance of the eight sentiment dictionaries on test data (80,000 book reviews). $Recall_{POS}$, $Recall_{NEG}$, and $Acc_{BAL}$ each indicate the classification accuracy of positive, negative, and overall reviews. Here, General Inquirer showed the best overall performance ($Acc_{BAL}$=65.8%); Opinion Lexicon and SentiWordNet performed well on positive reviews ($Recall_{POS}$=67.0%) whereas Subjectivity Lexicon performed well on negative reviews ($Recall_{NEG}$=70.6%).

Despite the significant difference between the General Inquirer and SentiWordNet's book review-related dictionary entry word sizes (Table

2: 3,853 vs. 33,923), the two exhibited comparable classification accuracies. The same can be said for the rest of the dictionaries excluding the lowest performing two dictionaries, Micro-WNOp and WordNet-Affect.

## 4 Combining Multiple Dictionaries

We now investigate the performance of combining multiple dictionaries through averaging, weighted-averaging, and majority voting.

### 4.1 McNemar's Test

We applied McNemar's test (McNemar, 1947) on the classification results of the individual sentiment dictionaries to investigate whether any two dictionaries' hits and misses were significantly different. The worst performing WordNet-Affect was excluded from the test.

Twenty-one dictionary pairs were generated from seven sentiment dictionaries. All dictionary pairs except the Opinion Lexicon vs. SentiWordNet (p=0.5552) exhibited significant differences in the proportion of hits and misses at 5% significance level[1].

### 4.2 Averaging, Weighted-Averaging, & Majority Voting

**Averaging:** The seven dictionaries' *RSS*s were averaged for each book review to calculate the combined *Averaged Review Sentiment Score* ($RSS_{AVG}$). We excluded the worst performing WordNet-Affect with lower than 50% accuracy since classifiers involved should provide a lower error rate than a random classifier (Enríquez *et al*., 2013).

$D_j$ indicates the individual dictionary, $j$ denotes the index of the sentiment dictionary, and $m$ indicates the number of sentiment dictionaries to be combined; in our case $m$ equals seven.

$$RSS_{AVG} = \frac{1}{m}\sum_{j=1}^{m} RSS(D_j)$$

$$ReviewLabel = \begin{cases} \text{positive} & \text{if } RSS_{AVG} \geq thres \\ \text{negative} & \text{otherwise} \end{cases}$$

A review was judged as positive if the $RSS_{AVG}$ was greater than or equal to the threshold, and as negative, otherwise. The threshold was determined using the threshold setting data.

[1] AFINN vs. SentiSense (p=5.221e-09), AFINN vs. Subjectivity Lexicon (p=0.0002395), General Inquirer vs. SentiSense (p=2.886e-12), and the remaining seventeen dictionary pairs (p<2.2e-16).

| | $Recall_{POS}$ | $Recall_{NEG}$ | $Acc_{BAL}$ | Thres |
|---|---|---|---|---|
| AVG | 66.7% | 68.9% | 67.8% | 0.115 |
| w-AVG | **67.3%** | 68.0% | 67.7% | 0.045 |
| Vote | 64.1% | **72.5%** | **68.3%** | N/A |

Table 5. Classification accuracy of the combined dictionaries using averaging (*AVG*), weighted-averaging (*w-AVG*), and majority voting (*Vote*).

**Weighted-Averaging:** The seven dictionaries' *RSS*s were weighted and averaged to yield a combined *Weighted-Averaged Review Sentiment Score* ($RSS_{w\text{-}AVG}$).

$$RSS_{w-AVG}(weight_j) = \frac{1}{m}\sum_{j=1}^{m} weight_j \cdot RSS(D_j)$$

$$\sum_j weight_j = 1, \ (0 \leq weight_j \leq 1)$$

$$ReviewLabel = \begin{cases} \text{positive} & \text{if } RSS_{w-AVG} \geq thres \\ \text{negative} & \text{otherwise} \end{cases}$$

Grid search was performed to set the weights of the seven dictionaries during the threshold setting stage. In the experiment, AFINN was given the greatest weight of 0.4, while the remaining six dictionaries were each given 0.1 weights.

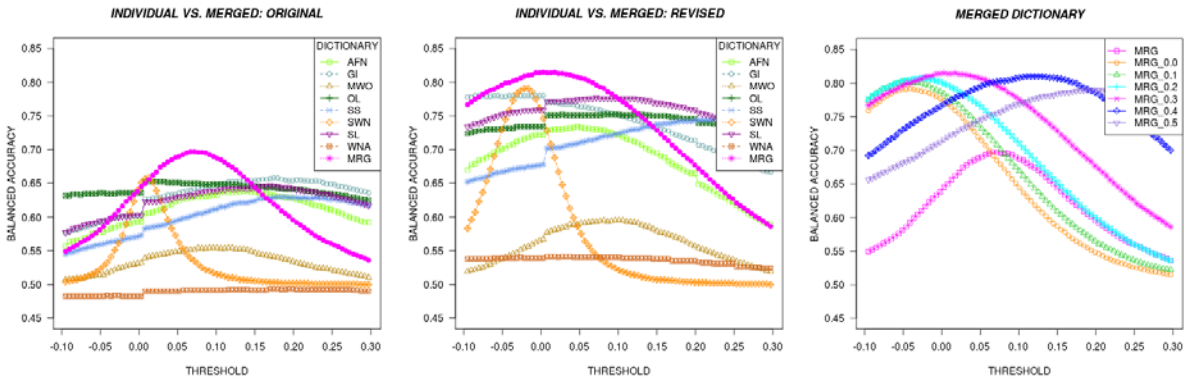
**Majority Voting**: The classification result (label) of each sentiment dictionary was used as votes in the majority voting. In the case of voting, the threshold was not set.

$$ReviewLabel = \begin{cases} \text{positive} & \text{if } Vote_{pos} > Vote_{neg} \\ \text{negative} & \text{otherwise} \end{cases}$$

Table 5 compares the classification accuracy of the three combined dictionaries on test data. *AVG*, *w-AVG*, and *Vote* each indicate averaging, weighted-averaging, and majority voting. The majority voting showed the best performance on the negative (72.5%) and overall (68.3%) review classification while the weighted-averaging showed the best performance on the positive (67.3%) review classification. However, the performance increase of the combined method was marginal compared to the best performing single dictionary (General Inquirer's 65.8% vs. majority voting's 68.3%).

### 4.3 Merging, Removing, & Switching

Combining multiple dictionaries at the surface level did not bring much improvement. We decided to merge the dictionaries at a deeper level and revise the dictionary entry's sentiment scores to improve the classification performance.




**(a)** **(b)** **(c)**

Figure 1. Classification performance of (a) eight original individual dictionaries (*AFN~WNA*) and one merged dictionary (*MRG*) and (b) revised dictionaries across different thresholds. (c) Classification performance of original merged dictionary (*MRG*) and revised merged dictionaries with different values of 'remove & switch' (*MRG_0.0~MRG_0.5*).

We could have merged all eight dictionaries, but instead merged the seven dictionaries excluding the SentiWordNet; we guessed that adding the largest SentiWordNet would simply result in an expanded version of the SentiWordNet and similar performance to the SentiWordNet. When merging the dictionary entries, the sentiment scores of the overlapping entry words, disregarding the parts-of-speech, were averaged. As a result, a merged sentiment dictionary containing 12,114 word entries was created. Threshold was also set for the merged dictionary, and the 80,000 book reviews were classified.

Table 6 compares the classification performance of the individual (*AFN~WNA*) and merged dictionaries (*MRG*). The first column lists the dictionaries (see Table 2 bottom for the full names of the dictionaries.), the second column displays the performance of the original dictionaries, and the third column shows the performance of the revised dictionaries. We confirmed that the merged dictionary (*MRG*) showed better performance (69.5%) than both the individual dictionaries and the best performing combined dictionary using majority voting (68.3%). Figure 1 (a) compares the performance of the nine dictionaries across different thresholds. We see that the merged dictionary (pink curve) outperforms the rest between the 0.05~0.10 threshold ranges.

Still, the performance of the merged dictionary did not improve dramatically. Therefore, we contrived a way to update the merged dictionary's entries to enhance performance. To do this, we leveraged the skewed distribution of positive/negative reviews. The general idea is to selectively (1) remove dictionary entries and (2) switch the polarity of sentiment scores.

| Senti. Dict. | Original $Acc_{BAL}$ (thres) | Revised $Acc_{BAL}$ (thres) |
|---|---|---|
| **AFN** | 63.8% (0.140) | 73.2% (0.030) |
| **GI** | 65.8% (0.175) | 78.0% (-0.085) |
| **MWO** | 55.4% (0.125) | 58.9% (0.045) |
| **OL** | 65.2% (0.025) | 75.1% (0.015) |
| **SS** | 62.8% (0.225) | 72.0% (0.085) |
| **SWN** | 65.7% (0.005) | 78.8% (-0.030) |
| **SL** | 64.7% (0.170) | 77.2% (0.005) |
| **WNA** | 49.0% (0.005) | 54.0% (0.075) |
| **MRG** | **69.5% (0.060)** | **80.9% (-0.025)** |

Table 6. Classification performance of original and revised dictionaries and their thresholds.

To implement the first idea, we removed those dictionary entry words with positive/negative book review word occurrence ratios that are similar to that of positive/negative book review ratio itself. The selection of the word was determined using the threshold setting data. For example, if the word "interested" appeared in the positive and negative reviews 900 and 100 times respectively, and the positive/negative review ratio of the threshold setting data is 9:1, we removed the "interested" entry from the dictionary. Such entry words were considered as not contributing to the actual classification.

To implement the second idea, we switched the sign of the selected dictionary entry words' sentiment scores whose positive/negative word occurrence ratio and the positive/negative review ratio's difference yielded a value with the sign opposite of its sentiment scores. For example, if the word "horror" appeared 900 and 300 times in the positive/negative book reviews resulting in 9:3 word occurrence ratios and the review ratio itself is 9:1, we calculated the difference between

the word and review ratio as 9/3 – 9/1, which resulted in 9/2, a positive number. However, the sentiment dictionary originally lists the sentiment score of "horror" as negative, e.g., -0.858; hence the sign (polarity) of the entry word "horror" was switched to positive, e.g., 0.858.

Table 6 shows the classification performance of the revised dictionaries. We see that the performance increased for all original dictionaries after they were revised using the 'remove & switch' procedure. The merged and revised dictionary showed the best performance (80.9%). Figure 1 (b) shows the performance of the revised dictionaries across different thresholds. We see that our method works better with larger dictionaries than smaller dictionaries such as *MWO* and *WNA*. This may be natural since our method includes the 'remove' procedure.

How much dictionary contents to 'remove & switch' were determined using the threshold setting data by experimenting with different proportion values. Figure 1 (c) compares the merged dictionary in its original version (*MRG*) and the revised versions using different values for revising (*MRG_0.0~MRG_0.5*). In our experiment, the best performing merged and revised dictionary's 'remove & switch' value was determined as 0.3 (*MRG_0.3*).

Our approach employs the most basic sentiment score aggregation to perform classification; no negation handling or structural analysis of the sentences is conducted. Our focus is on revising the sentiment dictionary by utilizing multiple dictionaries. At the outset, we surmised that combining and revising multiple dictionaries will have the following effects: (1) the word coverage will broaden and different dictionaries will complement each other. (2) The sentiment scores will be updated to incorporate diverse measurements leading to less odd scores.

However, broader coverage did not necessarily guarantee better performance since irrelevant words often matched to generate noise. By incorporating the '*remove*' procedure, we aimed to remove noise. Examples of removed words in the book reviews dataset included "book", "interested", and "mystery". With regard to the assumption (2) above, we found that contextual adjustment of sentiment scores was necessary for the given domain. Consequently we proposed the '*switch*' procedure which switched the polarity of selected dictionary entries. Examples of the switched words included "conspiracy", "horror", and "tragic" which were changed to have positive polarity.

## 5 Related Work

We were motivated by Taboada *et al*.'s (2011) work on lexicon-based sentiment analysis which couples hand-crafted sentiment dictionary with detailed sentence analysis. Although their sentiment calculation (SO-CAL) is more advanced than ours (it incorporates, for example, negation and intensification), we were able to confirm through the '*remove*' procedure that "less is more", i.e., less confounding dictionary entries will lead to more (greater) performance, with regard to the treatment of dictionary (Taboada *et al*., 2011; p.297). Our contribution is that we provided a simple data-based method to achieve "less is more" by leveraging the skewed distribution of the threshold setting positive/negative review data. This will be useful when ample threshold setting data is available, but dictionary expert is absent or costly.

Fahrni and Klenner (2008) proposed domain-specific adaptation of sentiment-bearing adjectives. Adjectives (e.g., good, bad, etc.) possess prior polarity, but depending on the context this polarity may change; for instance, warm mittens may be desirable, but warm beer may not be. To tackle the problem of contextual polarity, Fahrni and Klenner implemented a two-stage process that first identifies domain-specific targets using Wikipedia, and then determines the target-specific polarity of adjectives using a corpus. We performed a crude polarity adaptation by selectively switching the polarity of the dictionary entry's sentiment score based on the positive/negative distribution of the threshold setting data. Our approach, albeit crude, takes into account all dictionary entries, not restricted to adjectives, as candidates for polarity adaptation.

Neviarouskaya *et al*. (2011) described methods for automatically building and scoring new words based on sentiment-scored lemmas and types of affixes to create a sophisticated sentiment dictionary. Although we did not build sentiment dictionary from scratch, we experimented with shallow combinations and entry word merging of multiple dictionaries to show that shallow combination is insufficient, and that deeper-level merging and revising could be used as a viable method for enhancing the dictionary; in the process we generated revised dictionaries.

Various sentiment resources are built to perform different sentiment analysis tasks, so uniformly standardizing each resource may be unjust for some resources; moreover, we restrict our method's effectives within the sentiment

analysis of product reviews which is considered to be an easier problem compared to shorter texts such as microblogs (Cambria *et al*., 2013); we acknowledge these as our limitations.

## 6 Conclusion

We presented a method of merging multiple dictionaries, and removing and switching the merged dictionary's contents to achieve greater accuracy in the lexicon-based book review classification. In the future, we plan to investigate whether our approach is robust across different domains, how much threshold setting data is needed to achieve improvement in the revised dictionary, and what effects different positive/negative data distribution has on our method. We also plan to cover other sentiment resources such as SenticNet (Cambria *et al*., 2010) in the future.

**Acknowledgments**
This research was supported by the Korean Ministry of Science, ICT and Future Planning (MSIP) under the "IT Consilience Creative Program" supervised by the National IT Industry Promotion Agency (NIPA) of Republic of Korea. (NIPA-2013-H0203-13-1002)

## References

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the International Conference on Language Resources and Evaluation*.

Erik Cambria, Robert Speer, Catherine Havasi, and Amir Hussain. 2010. SenticNet: A publicly available semantic resource for opinion mining. In *Proceedings of the Association for the Advancement of Artificial Intelligence (AAAI) Fall Symposium*.

Erik Cambria, Björn Schuller, Yungqing Xia, and Catherine Havasi. 2013. New avenues in opinion mining and sentiment analysis. *IEEE Intelligent Systems*, 28(2):15-21.

Jorge Carrillo de Albornoz, Laura Plaza, and Pablo Gervás. 2012. SentiSense: An easily scalable concept-based affective lexicon for Sentiment Analysis. In *Proceedings of the International Conference on Language Resources and Evaluation*.

Fernando Enríquez, Fermín L. Cruz, F. Javier Ortega, Carlos, G. Vallejo, and José A. Troyano. 2013. A comparative study of classifier combination applied to NLP tasks. *Information Fusion*, 14(3):255-267.

Angela Fahrni and Manfred Klenner. 2008. Old wine or warm beer: Target-specific sentiment analysis of adjectives. In *Proceedings of Symposium on Affective Language in Human and Machine, AISB 2008 Convention*.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*.

Quinn McNemar. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153-157.

Alena Neviarouskaya, Helmut Prendinger, and Mitsuru Ishizuka. 2011. SentiFul: A lexicon for sentiment analysis. *IEEE Transactions on Affective Computing*, 2(1):22-36.

Finn Årup Nielsen. 2011. A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. In *Proceedings of the ESWC Workshop on Making Sense of Microposts*.

Ellen Riloff and Janyce Wiebe. 2003. Learning extraction patterns for subjective expressions. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Philip J. Stone and Earl B. Hunt. 1963. A computer approach to content analysis: Studies using the General Inquirer system. In *Proceedings of the American Federation of Information Processing Societies, Spring Joint Computer Conference*.

Carlo Strapparava and Alessandro Valitutti. 2004. WordNet-Affect: An affective extension of WordNet. In *Proceedings of the International Conference on Language Resources and Evaluation*.

Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. 2011. Lexicon-based methods for sentiment analysis. *Computational Linguistics*, 37(2): 267-307.

Kristina Toutanova, Dan Klein, Christopher D. Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*.