Experimental Evaluation of a Lexicon- and Corpus-based Ensemble for Multi-way Sentiment Analysis*

Minh Duc Cao
School of Chemistry and Molecular Biosciences
The University of Queensland
St Lucia, QLD 4072, Australia
m.cao1@uq.edu.au

Ingrid Zukerman
Clayton School of Information Technology
Monash University
Clayton, VIC 3800, Australia
Ingrid.Zukerman@monash.edu

Abstract

We describe a probabilistic approach that combines information obtained from a lexicon with information obtained from a Naïve Bayes (NB) classifier for multi-way sentiment analysis. Our approach also employs grammatical structures to perform adjustments for negations, modifiers and sentence connectives. The performance of this method is compared with that of an NB classifier with feature selection, and MCST – a state-of-the-art system. The results of our evaluation show that the performance of our hybrid approach is at least as good as that of these systems. We also examine the influence of three factors on performance: (1) sentiment-ambiguous sentences, (2) probability of the most probable star rating, and (3) coverage of the lexicon and the NB classifier. Our results indicate that the consideration of these factors supports the identification of regions of improved reliability for sentiment analysis.

1 Introduction

A key problem in sentiment analysis is to determine the polarity of sentiment in text. Much of the work on this problem has considered binary sentiment polarity (positive or negative) at granularity levels ranging from sentences (Mao and Lebanon, 2006; McDonald et al., 2007) to documents (Wilson et al., 2005; Allison, 2008). Multi-way polarity classification, i.e., the problem of inferring the "star" rating associated with a review, has been attempted in several domains, e.g., restaurant reviews (Snyder and Barzilay, 2007) and movie reviews (Bickerstaffe and Zukerman, 2010; Pang and Lee, 2005). Star ratings are more informative than positive/negative ratings, and are commonly given in reviews of films, restaurants, books and consumer goods. However, because of this finer grain, multi-way sentiment classification is a more difficult task than binary classification. Hence, the results for multi-way classification are typically inferior to those obtained for the binary case.

Most of the research on sentiment analysis uses supervised classification methods such as Maximum Entropy (Berger et al., 1996), Support Vector Machines (SVMs) (Cortes and Vapnik, 1995) or Naïve Bayes (NB) (Domingos and Pazzani, 1997). The sentiment expressed in word patterns has been exploited by considering word n-grams (Hu et al., 2007), applying feature selection to handle the resultant proliferation of features (Mukras et al., 2007). In addition, when performing multi-way classification, approaches that consider class-label similarities (Bickerstaffe and Zukerman, 2010; Pang and Lee, 2005) generally outperform those that do not.

Lexicon-based methods for sentiment analysis have been investigated in (Beineke et al., 2004; Taboada et al., 2011; Andreevskaia and Bergler, 2008; Melville et al., 2009) in the context of binary, rather than multi-way, sentiment classifiers. These methods often require intensive labour (e.g., via the Mechanical Turk service) to build up the lexicon (Taboada et al., 2011) or use a small, generic lexicon enhanced by sources from the Internet (Beineke et al., 2004). Andreevskaia and Bergler (2008) and Melville et al. (2009) employ a weighted average to combine information from the lexicon with the classifi-

*The majority of this work was done while the first author was at Monash University.

Minh Duc Cao and Ingrid Zukerman. 2012. Experimental Evaluation of a Lexicon- and Corpus-based Ensemble for Multi-way Sentiment Analysis. In *Proceedings of Australasian Language Technology Association Workshop*, pages 52−60.

cation produced by a supervised machine learning method. Their results demonstrate the effectiveness of these methods only on small datasets, where the contribution of the machine-learning component is limited.

This paper examines the performance of a hybrid lexicon/supervised-learning approach and two supervised machine learning methods in multi-way sentiment analysis. The hybrid approach combines information obtained from the lexicon with information obtained from an NB classifier with feature selection. Information is obtained from a lexicon by means of a novel function based on the Beta distribution. This function, which employs heuristics to account for negations, adverbial modifiers and sentence connectives, combines the sentiment of words into the sentiment of phrases, sentences, and eventually an entire review (Section 2). The supervised learning methods are: an NB classifier with feature selection, and MCST (Bickerstaffe and Zukerman, 2010) – a state-of-the-art classifier based on hierarchical SVMs which considers label similarity (MCST outperforms Pang and Lee's (2005) best-performing methods on the Movies dataset described in Section 5.1).

We also investigate the influence of three factors on sentiment-classification performance: (1) presence of sentiment-ambiguous sentences, which we identify by means of a heuristic (Section 4); (2) probability of the most probable star rating; and (3) coverage of the lexicon and the NB classifier, i.e., fraction of words in a review being "understood".

Our results show that (1) the hybrid approach generally performs at least as well as NB with feature selection and MCST; (2) NB with feature selection generally outperforms MCST, highlighting the importance of choosing stringent baselines in algorithm evaluation; (3) the performance of sentiment analysis algorithms deteriorates as the number of sentiment-ambiguous sentences in a review increases, and improves as the probability of the most probable star rating of a review increases (beyond 50%), and as the coverage of the lexicon and the NB classifier increases (between 50% and 80%).

In the next section, we present our lexicon-based approach. Section 3 describes the combination of the lexicon with an NB classifier, followed by our heuristic for identifying sentiment-ambiguous sentences. Section 5 presents the results of our evaluation, and Section 6 offers concluding remarks.

2 Harnessing the Lexicon

In this section, we present our framework for representing information from a lexicon, and combining this information into phrases, sentences and entire reviews, and our heuristics for modifying the sentiment of a word or phrase based on grammatical information. We report on the results obtained with the lexicon collected by Wilson et al. (2005), which contains 8221 sentiment-carrying words (most are open-class words, but there are a few modals, conjunctions and prepositions); each word is identified as positive, negative or neutral, and either strong or weak.[1]

The numeric rating of a review is inferred from the sentiment of the words in it, while taking into account the uncertainty arising from (1) the ambiguous sentiment of individual words, and (2) our ignorance due to the lack of understanding of the sentiment of some words. Instead of committing to a particular star rating for a review, we assign a probability to each star rating and return the most probable star rating. This probability is modelled by a *unimodal* distribution, as the rating of a review is likely to be centered around the most probable star rating. For example, if a review is most likely to be in the 4-star class, the probability of this review having 3 stars should be higher than the probability of 2 stars.

We chose the Beta distribution to represent sentiment information because (1) its parameters α and β, which encode the *positiveness* and *negativeness* of the distribution respectively, are well-suited to represent the sentiment of every linguistic entity (i.e., word, phrase, sentence or review); and (2) it has appealing computational properties which facilitate the combination of the Beta distributions of those entities. The combination of the distributions of the words in a sentence yield a Beta distribution for the sentence, and the combination of the distributions for the sentences in a review yield a Beta distribution for the review.

To fully exploit the grammatical structure of a sentence, we first parse the sentence using the Stanford parser (Klein and Manning, 2003). We

[1] We also considered SentiWordNet (Baccianella et al., 2010), but it yielded inferior results.

then map the sentiment values of a word from the lexicon to the α and β parameters of the Beta distribution for the word, while maintaining the constraint $\alpha + \beta = 1$ (this constraint is relaxed for phrases, sentences and the entire review). Specifically, $\alpha = 1$ for a strong positive word, and $\beta = 1$ for a strong negative word; a weak positive word is assigned $\alpha = 0.75$, and a weak negative word $\beta = 0.75$; and $\alpha = \beta = 0.5$ for a neutral word.

We employ the function $\oplus$ to combine the distributions of individual words into distributions of successively higher-level segments in the parse tree, until we obtain the distribution of a whole sentence and then an entire review. For example, given review $R = \{(w_1 w_2).(w_3(w_4 w_5))\}$ comprising two main sentences $w_1 w_2$ and $w_3(w_4 w_5)$, its density function f is defined as $f(R) = (w_1 \oplus w_2) \oplus (w_3 \oplus (w_4 \oplus w_5))$. Unless otherwise specified, $\oplus$ multiplies the probabilities of consecutive segments. This is conveniently equivalent to adding the α and β values of the segments, i.e., $f(\alpha_1, \beta_1) f(\alpha_2, \beta_2) = f(\alpha_1 + \alpha_2, \beta_1 + \beta_2)$.

The probability that review R has a rating k is

$$\Pr(\textit{rating}(R) = k) = \int_{b_{k-1}}^{b_k} f(y)dy \qquad (1)$$

where b_i is the upper boundary of rating i ($0 = b_0 < b_1 < \ldots < b_N = 1$), and N is the highest star rating. These boundaries were determined by a hill-climbing algorithm that maximizes classification accuracy on the training set.

Special operators, such as negations, adverbial modifiers and sentence connectives, alter the definition of the $\oplus$ function as follows (our identification of negations and modifiers resembles that in (Taboada et al., 2011), but our mappings are probabilistic).

Negations. Negations often shift the sentiment of a word or a phrase in the opposite direction (rather than inverting its polarity), e.g., "not outstanding" is better than "not good" (Taboada et al., 2011). This idea is implemented by adjusting the α and β parameters so that the new parameters α' and β' obey the constraint $\alpha' + \beta' = \alpha + \beta$, and the new mean of the distribution is

$$\frac{\alpha'}{\alpha' + \beta'} = \frac{\alpha}{\alpha + \beta} + \lambda$$

where $\lambda = -0.5$ for positive words/phrases and $+0.5$ for negative ones. For instance, based on the lexicon, $\alpha_{\text{good}} = 0.75$ ($\beta_{\text{good}} = 0.25$), which yields $\alpha'_{\text{not good}} = 0.25$ ($\beta'_{\text{not good}} = 0.75$). This procedure is also applied to antonyms of words in the lexicon, which are identified by removing a negation prefix from an input word (e.g., *un-, in-, il-, im-, de-, ab-, non-, dis-*), and matching with the lexicon, e.g., "*un*able" shifts the sentiment of "able". The combination of a negation and a phrase, e.g., "I *don't* think (the staff is friendly and efficient enough)", has the same effect.

Table 1: Sample modifications of the word *polite* ($\alpha = 0.75$ and $\beta = 0.25$)

Adverb	γ	α'	β'
hardly	-0.9	0.525	0.475
relatively	-0.1	0.725	0.275
more	0.4	0.850	0.150
really	0.7	0.925	0.075
absurdly	0.8	0.950	0.050
completely	1.0	1.000	0

Adverbial modifiers. Adverbs normally change the intensity of adjectives or verbs (e.g., "very" is an intensifier, while "hardly" is a diminisher). Like Taboada et al. (2011), we increase or decrease the sentiment level of a word based on the meaning of its modifier. This is done by adjusting the α' and β' of weak adjectives and verbs as follows (currently, we leave strong words unchanged as it is unusual to apply adverbial modifiers to such words, e.g., "somewhat excellent"): $\alpha' = \alpha \pm \gamma\beta$ and $\beta' = \beta \mp \gamma\beta$, where the sign is determined by the polarity of the word, and γ is determined by the adverb. For example, $\gamma = -0.2$ for "fairly" and $\gamma = 0.5$ for "very". Thus, "fairly polite" moves "polite" from $\alpha = 0.75$ ($\beta = 0.25$) to $\alpha = 0.7$ ($\beta = 0.3$). Table 1 shows the intensity level γ of several adverbs, and their effect on the polarity of the adjective "polite".

Dealing with uncertainty. When reading a text, the number of words a reader understands affects his/her confidence in his/her comprehension. The fewer words are understood, the higher the reader's uncertainty. We estimate w_i, the level of comprehension of sentence s_i, by means of the fraction of open-class and lexicon words in the sentence that appear in the lexicon (recall that the lexicon contains some closed-class words). When combining the sentiment derived from two sentences s_1 and s_2, we want the sentence that is less

understood to carry a lower weight than the sentence that is better understood. To implement this idea, we adjust the probability of the star rating of a sentence by a function of the certainty of understanding it. We employ an exponential function as follows, where the exponents are the above weights w_i.

$$\Pr(y|s_1, s_2) \propto \Pr(y|s_1)^{w_1} \Pr(y|s_2)^{w_2} \quad (2)$$

Since $0 \leq w_i \leq 1$, a low certainty for w_i yields a value close to 1, which has relatively little effect on the outcome, while a high certainty has a large effect on the outcome.

Sentence connectives. When we have little confidence in our understanding of a sentence, sentence connectives, such as adversatives (e.g., "but", "however") or intensifiers (e.g., "furthermore"), may prove helpful. Assume that sentence s_1 has an adversative relation with sentence s_2, and w.l.o.g., assume that s_1 is better understood than s_2 (i.e., $w_1 > w_2$, where w_i is the level of comprehension of sentence s_i). We model the idea that in this case, the sentiment of s_2 is likely to contradict that of s_1 by shifting the sentiment of s_2 closer to that of $\bar{s_1}$ (the negation of s_1) in proportion to the difference between the weights of these sentences.

$$\Pr'(y|s_2) = \frac{\Pr(y|s_2)w_2 + \Pr(y|\bar{s_1})(w_1 - w_2)}{w_1} \quad (3)$$

In addition, owing to the interaction between s_2 and s_1, w_2 increases to $w_2' = \frac{1}{2}(w_1 + w_2)$ to indicate that s_2 is now better understood. For example, consider a situation where the probability that sentence s_1 conveys a 4-star rating is 0.2 with $w_1 = 0.8$ (four fifths of the words in s_1 were understood), and the probability that s_2 conveys a 4-star rating is 0.4 with $w_2 = 0.2$. Further, assume that there is an adversative relation between s_1 and s_2, e.g., "s_1. However, s_2". After applying Equation 3 to adjust the probability of the less understood sentence, s_2, we obtain $\Pr'(y = 4\,\text{stars}|s2) = (0.4 \times 0.2 + 0.6\,(0.8 - 0.2))/0.8 = 0.55$, and $w_2' = 0.5$ (the 0.6 is obtained by negating s_1). Thus, the probability that s_2 conveys a 4-star rating has increased, as has the certainty of this assessment.

Parameterization and heuristics. The values of the different parameters $(\alpha, \beta, \gamma, \delta, \lambda)$ were manually determined. We tried several combinations, but the effect was negligible, arguably due to the low coverage of the lexicon (Section 5). Further, we employ different types of heuristics, e.g., the modification of the probabilities of individual sentences is additive, while sentence combination is multiplicative (as per the Beta distribution). The application of machine learning techniques or a hill-climbing procedure to determine parameter values that yield improved performance, as well as the consideration of different heuristics for negations, adverbial modifiers, sentence connectives and dealing with uncertainty, may be a profitable avenue of investigation after lexicon coverage is increased.

3 Combining the Lexicon with a Naïve Bayes Classifier

Beineke et al. (2004) combined a lexicon with an NB classifier by sourcing from a large corpus words that co-occur with known sentimental "anchor" words, and employing these words to train the classifier. In contrast, like Andreevskaia and Bergler (2008) and Melville et al. (2009), we combine information from a lexicon with the classification produced by a supervised machine learning method. However, in their systems, the weights assigned to each contributing method are based on this method's performance on the training set, while our weights represent a method's coverage of the current text. In addition, we employ much larger datasets in our experiments than those used in (Andreevskaia and Bergler, 2008) and (Melville et al., 2009), and unlike them, we take into account negations, adverbial modifiers and sentence connectives to modify the sentiment of lexicon words.

Our system incorporates corpus-based information by training an NB classifier with unigrams and bigrams as features, and applying information gain (Yang and Pedersen, 1997) to select the top K $(= 4000)$ features.[2] This version of NB is denoted **NB4000**. The probability obtained from the classifier for a review is combined with that obtained from the lexicon by means of a weighted average.[3]

[2] According to our experiments, NB classifiers trained using unigrams and bigrams, combined with feature selection, are among the best sentiment classifiers.

[3] We also applied this combination procedure at the sentence level, but with inferior results.

$$\Pr_{COMB}(D|s) = \frac{\Pr_{NB}(D|s)w_{NB}+\Pr_{LEX}(D|s)w_{LEX}}{w_{NB}+w_{LEX}} \quad (4)$$

where D is a document; w_{LEX} is the fraction of open-class and lexicon words in the review that appear in the lexicon; and w_{NB} represents the fraction of *all* the words in the review that appear in the NB features (this is because unigrams and bigrams selected as NB features contain both open- and closed-class words).

4 Identifying Bimodal Sentences

Sentiment analysis is a difficult problem, as opinions are often expressed in subtle ways, such as irony and sarcasm (Pang and Lee, 2008), which may confuse human readers, creating uncertainty over their understanding. In Section 2, we discussed the incorporation of uncertainty into the lexicon-based framework. Here we offer a method for identifying reviews that contain sentiment-ambiguous sentences, which also affect the ability to understand a review.

As mentioned above, the probability distribution of the sentiment in a review is likely to be unimodal. The Beta distribution obtained from the lexicon guarantees this property, but the multinomial distribution used to train the NB classifier does not. Further, the combination of the distributions obtained from the lexicon and the NB classifier can lead to a bimodal distribution due to inconsistencies between the two input distributions. We posit that such bimodal sentences are unreliable, and propose the following heuristic to identify bimodal sentences.[4]

> The sentiment distribution in a sentence is bimodal if (1) the two most probable classes are not adjacent (e.g., 2-star and 4-star rating), and (2) the probability of the second most probable class is more than half of that of the most probable class.

Examples of sentences identified by this heuristic are "It is pretty *boring*, but you do not worry because the picture will be *beautiful*, and you have these *gorgeous* stars too" (NB⇒1, Lexicon⇒3, actual = 1) and " 'The Wonderful, Horrible Life of Leni Riefenstahl' is a *excellent* film, *but it needed* Riefenstahl to edit it more" (NB⇒2&4,Lexicon⇒3, actual=4). The impact of bimodal sentences on performance is examined in Section 5.2.

[4] A statistical method for identifying bi-modality is described in (Jackson et al., 1989).

5 Evaluation

5.1 Datasets

Our datasets were sourced from reviews in various domains: movies, kitchen appliances, music, and post office. These datasets differ in review length, word usage and writing style.

- **Movies**[5]: This is the *Sentiment Scale* dataset collected and pre-processed by Pang and Lee (2005), which contains movie reviews collected from the Internet. They separated the dataset into four sub-corpora, each written by a different author, to avoid the need to calibrate the ratings given by different authors. The authors, denoted A, B, C and D, wrote 1770, 902, 1307 and 1027 reviews respectively. Each author's reviews were grouped into three and four classes, denoted **Author*X*3** and **Author*X*4** respectively, where $X \in \{A, B, C, D\}$.

- **Kitchen**[6]: This dataset was sourced from a large collection of kitchen appliance reviews collected by Blitzer et al. (2007) from Amazon product reviews. We selected 1000 reviews from each of the four classes considered by Blitzer *et al.*, totalling 4000 reviews. The resultant dataset is denoted **Kitchen4**.

- **Music**[7]: We selected 4039 text samples of music reviews from the Amazon product review dataset compiled by Jindal and Liu (2008). To obtain a dataset with some degree of item consistency and reviewer reliability, we selected reviews for items that have at least 10 reviews written by users who have authored at least 10 reviews. The original reviews are associated with a 5-point rating scale, but we grouped the reviews with low ratings ($\leq$ 3 stars) into one class due to their low numbers. The resultant dataset, denoted

[5] http://www.cs.cornell.edu/home/llee/data/

[6] http://www.cs.jhu.edu/~mdredze/datasets/sentiment/

[7] http://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html

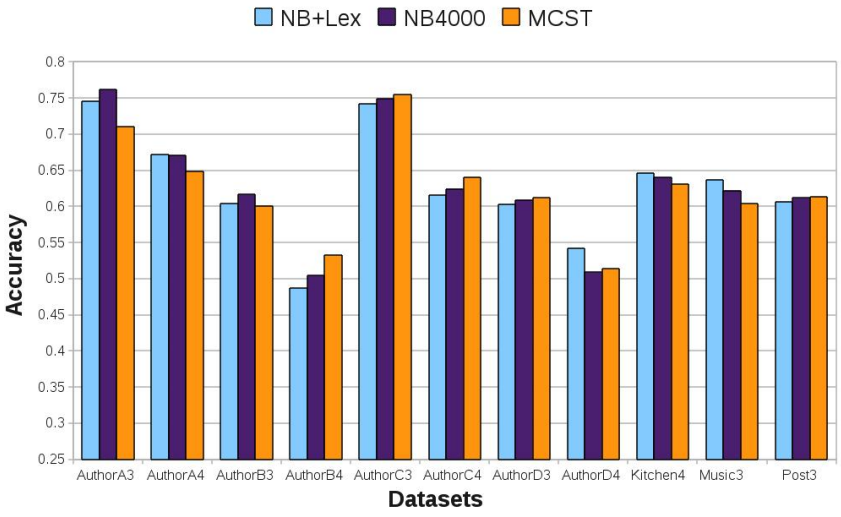


Figure 1: Average classification accuracy for NB+Lex, compared with NB4000 and MCST; all datasets.

Music3, contains three classes: 860 low reviews ($\leq$ 3 stars), 1409 medium (4 stars) and 1770 high (5 stars).

- **PostOffice**: contains 3966 reviews of post-office outlets written by "mystery shoppers" hired by a contractor. The reviews are very short, typically comprising one to three sentences, and focus on specific aspects of the service, e.g., attitude of the staff and cleanliness of the stores. The reviews were originally associated with a seven-point rating scale. However, as for the Music dataset, owing to the low numbers of reviews with low ratings ($\leq$ 5 stars), we grouped the reviews into three balanced classes denoted **Post3**: 1277 low reviews ($\leq$ 5 stars), 1267 medium (6 stars), and 1422 high (7 stars).

5.2 Results

Figure 1 shows the average accuracy obtained by the hybrid approach (NB+Lex using NB4000),[8] compared with the accuracy obtained by the best-performing version of MCST (Bickerstaffe and Zukerman, 2010) (which was evaluated on the Movies dataset, using the algorithms presented in (Pang and Lee, 2005) as baselines), and by NB4000 (NB plus feature selection with 4000 features selected using information gain). All trials employed 10-fold cross-validation. For the NB+Lex method, we investigated different combinations of settings (with and without negations, modifiers, sentence connectives, and inter-sentence weighting). However, these variations had a marginal effect on performance, arguably owing to the low coverage of the lexicon. Here we report on the results obtained with all the options turned on. Statistical significance was computed using a two-tailed paired t-test for each fold with $p = 0.05$ (we mention only statistically significant differences in our discussion).

- NB+Lex outperforms MCST on three datasets (AuthorA3, AuthorA4 and Music3), while the inverse happens only for AuthorB4. NB+Lex also outperforms NB4000 on AuthorD4 and Music3. No other differences are statistically significant.

- Interestingly, NB4000 outperforms MCST for AuthorA3 and Music3, with no other statistically significant differences, which highlights the importance of judicious baseline selection.

Despite showing some promise, it is somewhat disappointing that the combination approach does not yield significantly better results than NB4000 for all the datasets. The small contribution of the lexicon to the performance of NB+Lex may be partially attributed to its low coverage of the vocabulary of the datasets compared with the coverage of NB4000 alone. Specifically, only a small

[8] We omit the results obtained with the lexicon alone, as its coverage is too low.

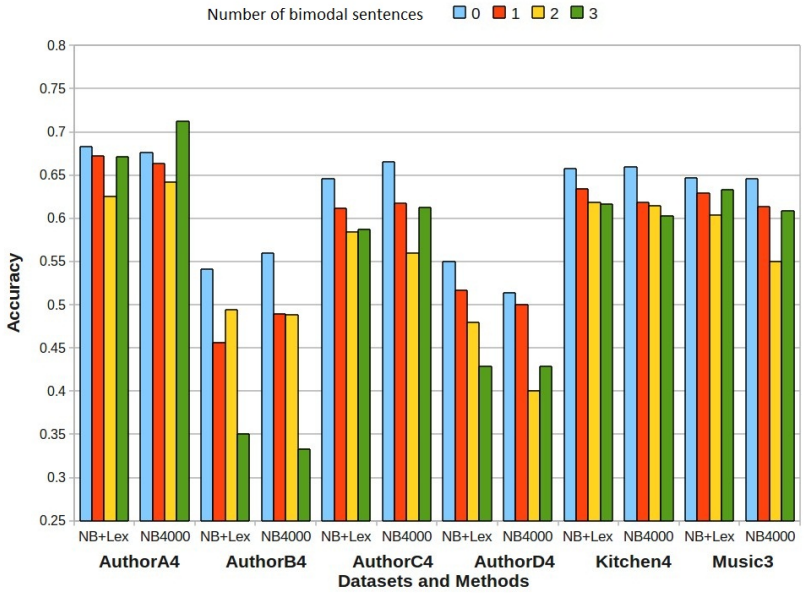


Figure 2: Effect of bimodal sentences on performance (average accuracy): NB+Lex and NB4000; datasets AuthorA4-D4, Kitchen4, Music3.

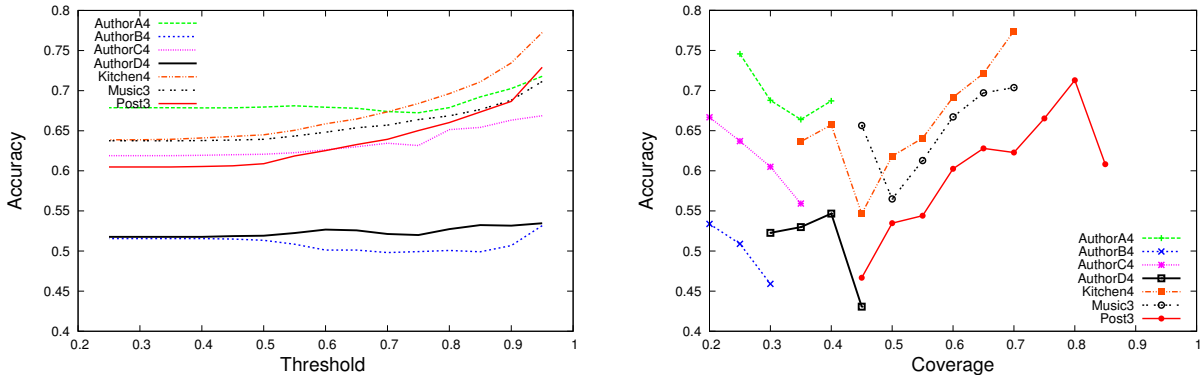
fraction of the words in our datasets is covered by the lexicon (between 5.5% for AuthorC and 7% for PostOffice), compared with the NB4000 coverage (between 31% for AuthorA3 and 67% for PostOffice). Further, as indicated above, the formulas for estimating the influence of negations, adverbial modifiers and sentence connectives are rather *ad hoc*, and the parameters were manually set. Different heuristics and statistical methods to set their parameters warrant future investigation.

It is interesting to note that the overlap between the words in the corpora covered by the lexicon and the words covered by the 4000 features used by NB is rather small. Specifically, in all datasets except PostOffice, which has an unusually small vocabulary (less than 3000 words), between half and two thirds of the lexicon-covered words are *not* covered by the set of 4000 features. This discrepancy in coverage means that the unigrams in the lexicon have a lower information gain, and hence are less discriminative, than many of the 4000 features selected for the NB classifier, which include a large number of bigrams.

We also analyzed the effect of the presence of sentiment-ambiguous (bimodal) sentences on the predictability of a review, using the method described in Section 4 to identify bimodal sentences. Figure 2 displays the accuracy obtained by NB+Lex and NB4000 on the datasets Authors4A-D, Kitchen4 and Music3 as a function of the number of bimodal sentences in a review (the Authors3A-D datasets were omitted, as they are "easier" than Authors4A-D, and Post3 was omitted because of the low number of sentences per review). We display only results for reviews with 0 to 3 bimodal sentences, as there were very few reviews with more bimodal sentences. As expected, performance was substantially better for reviews with no bimodal sentences (with the exception of NB4000 on AuthorsA4 with 3 bimodal sentences per review). These results suggest that the identification of bimodal sentences is worth pursuing, possibly in combination with additional lexicon coverage, to discriminate between reviews whose sentiment can be reliably detected and reviews where this is not the case. Further, it would be interesting to ascertain the views of human annotators with respect to the sentences we identify as bimodal.

In the context of identifying difficult reviews, we also investigated the relationship between prediction confidence (the probability of the most probable star rating in a review) and performance (Figure 3(a)), and between the coverage provided by both the lexicon and the NB classifier and performance (Figure 3(b)[9]). As seen in Figure 3(a), for all datasets, except AuthorB4, accuracy improves as prediction confidence increases. This

[9]We do not display results for less than 50 documents with a particular coverage.



(a) Probability of the most probable star rating.

(b) Lexicon and NB coverage.

Figure 3: Relationship between probability of the most probable star rating and accuracy, and between lexicon/NB coverage and accuracy; datasets AuthorsA4-D4, Kitchen4, Music3 and Post3.

improvement is steadier and sharper for Kitchen4, Music3 and Post3, which as seen in Figure 3(b), have a higher lexicon and NB coverage than the Authors datasets. As one would expect, performance improves for the first three datasets as coverage increases from 50% to 80%. However, outside this range, the results are counter-intuitive: overall, accuracy decreases between 20% and 50% coverage, and also drops for Post3 at 85% coverage (a level of coverage that is not obtained for any other dataset); and a high level of accuracy is obtained for very low levels of coverage ($\leq 25\%$) for AuthorA4 and AuthorC4. These observations indicate that other factors, such as style and vocabulary, should be considered in conjunction with coverage, and that the use of coverage in Equations 2 and 4 may require fine-tuning to take into account the level of coverage.

6 Conclusion

We have examined the performance of three methods based on supervised machine learning applied to multi-way sentiment analysis: (1) sentiment lexicon combined with NB with feature selection, (2) NB with feature selection, and (3) MCST (which considers label similarity). The lexicon is harnessed by applying a probabilistic procedure that combines words, phrases and sentences, and performs adjustments for negations, modifiers and sentence connectives. This information is combined with corpus-based information by taking into account the uncertainty arising from the extent to which the text is "understood".

Our methods were evaluated on seven datasets of different sizes, review lengths and writing styles. The results of our evaluation show that the combination of lexicon- and corpus-based information performs at least as well as state-of-the-art systems. The fact that this improvement is achieved with a small contribution from the lexicon indicates that there may be promise in increasing lexicon coverage and improving the domain specificity of lexicons. At the same time, the observation that NB+Lex (with a small lexicon), NB4000 and MCST exhibit similar performance for several datasets leads us to posit that pure n-gram based statistical systems have plateaued, thus reinforcing the point that additional factors must be brought to bear to achieve significant performance improvements.

The negative result that an NB classifier with feature selection achieves state-of-the-art performance indicates that careful baseline selection is warranted when evaluating new algorithms.

Finally, we studied the effect of three factors on the reliability of a sentiment-analysis algorithm: (1) number of bimodal sentences in a review; (2) probability of the most probable star rating; and (3) coverage provided by the lexicon and the NB classifier. Our results show that these factors may be used to predict regions of reliable sentiment-analysis performance, but further investigation is required regarding the interaction between coverage and the stylistic features of a review.

References

B. Allison. 2008. Sentiment detection using lexically-based classifiers. In *Proceedings of the 11th International Conference on Text, Speech and Dialogue*, pages 21–28, Brno, Czech Republic.

A. Andreevskaia and S. Bergler. 2008. When specialists and generalists work together: Overcoming domain dependence in sentiment tagging. In *ACL'08 – Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics*, pages 290–298, Columbus, Ohio.

S. Baccianella, A. Esuli, and F. Sebastiani. 2010. SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, and D. Tapias, editors, *LREC'10 – Proceedings of the 7th Conference on International Language Resources and Evaluation*, Valletta, Malta.

P. Beineke, T. Hastie, and S. Vaithyanathan. 2004. The sentimental factor: Improving review classification via human-provided information. In *ACL'04 – Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, pages 263–270, Barcelona, Spain.

A.L. Berger, Stephen A. Della Pietra, and Vincent J. Della Pietra. 1996. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):39–71.

A. Bickerstaffe and I. Zukerman. 2010. A hierarchical classifier applied to multi-way sentiment detection. In *COLING'2010 – Proceedings of the 23rd International Conference on Computational Linguistics*, pages 62–70, Beijing, China.

J. Blitzer, M. Dredze, and F. Pereira. 2007. Biographies, Bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *ACL'07 – Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, pages 440–447, Prague, Czech Republic.

C. Cortes and V. Vapnik. 1995. Support Vector Networks. *Machine Learning*, 20(3):273–297.

P. Domingos and M. Pazzani. 1997. On the optimality of the simple Bayesian classifier under zero-one loss. *Machine Learning*, 29:103–130.

Y. Hu, R. Lu, X. Li, Y. Chen, and J. Duan. 2007. A language modeling approach to sentiment analysis. In *ICCS07 – Proceedings of the 7th International Conference on Computational Science, Part II*, pages 1186–1193, Beijing, China.

P.R. Jackson, G.T. Tucker, and H.F. Woods. 1989. Testing for bimodality in frequency distributions of data suggesting polymorphisms of drug metabolism-hypothesis testing. *British Journal of Clinical Pharmacology*, 28:655–662.

N. Jindal and B. Liu. 2008. Opinion spam and analysis. In *WSDM-2008 – Proceedings of the 1st International Conference on Web Search and Web Data Mining*, pages 219–230, Palo Alto, California.

D. Klein and C.D. Manning. 2003. Accurate unlexicalized parsing. In *ACL'03 – Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 423–430, Sapporo, Japan.

Y. Mao and G. Lebanon. 2006. Isotonic conditional random fields and local sentiment flow. In *NIPS 2006 – Proceedings of the 20th Annual Conference on Neural Information Processing Systems*, pages 961–968, British Columbia, Canada.

R. McDonald, K. Hannan, T. Neylon, M. Wells, and J. Reynar. 2007. Structured models for fine-to-coarse sentiment analysis. In *ACL'07 – Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, pages 432–439, Prague, Czech Republic.

P. Melville, W. Gryc, and R.D. Lawrence. 2009. Sentiment analysis of blogs by combining lexical knowledge with text classification. In *KDD'09 – Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1275–1284, Paris, France.

R. Mukras, N. Wiratunga, R. Lothian, S. Chakraborti, and D. Harper. 2007. Information gain feature selection for ordinal text classification using probability re-distribution. In *Proceedings of the IJCAI-07 TextLink Workshop*, Hyderabad, India.

B. Pang and L. Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *ACL'05 – Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, pages 115–124, Ann Arbor, Michigan.

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

B. Snyder and R. Barzilay. 2007. Multiple aspect ranking using the Good Grief algorithm. In *NAACL-HLT 2007 – Proceedings of Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 300–307, Rochester, New York.

M. Taboada, J. Brooke, M. Tofiloskiy, K. Vollz, and M. Stede. 2011. Lexicon-based methods for sentiment analysis. *Computational Linguistics*, 37(2):267–307.

T. Wilson, J. Wiebe, and P. Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *EMNLP'2005 – Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 347–354, Vancouver, Canada.

Y. Yang and J.O. Pedersen. 1997. A comparative study on feature selection in text categorization. In *ICML'97 – Proceedings of the 14th International Conference on Machine Learning*, pages 412–420, Nashville, Tennessee.