Khoo, C.S.G., Johnkhan, Sathik B., & Na, J.C. (2015). Evaluation of a general-purpose sentiment lexicon on a product review corpus. In R.B. Allen, J. Hunter, & M.L. Zeng (Eds.), 17th International Conference on Asia-Pacific Digital Libraries, ICADL2015: Proceedings (LNCS 9469, pp. 82–93). Berlin: Springer.

Evaluation of a General-Purpose Sentiment Lexicon on a Product Review Corpus

Christopher S.G. Khoo(✉), Sathik Basha Johnkhan, and Jin-Cheon Na

Wee Kim Wee School of Communication & Information,
Nanyang Technological University, Singapore, Singapore
chriskhoo@pmail.ntu.edu.sg, {sathik,TJCNa}@ntu.edu.sg

Abstract. This paper introduces a new general-purpose sentiment lexicon called the WKWSCI Sentiment Lexicon and compares it with three existing lexicons. The WKWSCI Sentiment Lexicon is based on the *6of12dict* lexicon, and currently covers adjectives, adverbs and verbs. The words were manually coded with a value on a 7-point sentiment strength scale. The effectiveness of the four sentiment lexicons for sentiment categorization at the document-level and sentence-level was evaluated using an Amazon product review dataset. The WKWSCI lexicon obtained the best results for document-level sentiment categorization, with an accuracy of 75%. The Hu & Liu lexicon obtained the best results for sentence-level sentiment categorization, with an accuracy of 77%. The best bag-of-words machine learning model obtained an accuracy of 82% for document-level sentiment categorization model. The strength of the lexicon-based method is in sentence-level and aspect-based sentiment analysis, where it is difficult to apply machine-learning because of the small number of features.

Keywords: Sentiment lexicon · Sentiment analysis · Sentiment categorization

1 Introduction

Digital libraries increasingly contain user-contributed content in the form of user comments and reviews on the digital library materials. Some digital libraries, especially those of cultural heritage materials, contain crowdsourced content [1]. User-contributed materials are more likely to contain subjective content and sentiment expressions. It will become desirable to be able to categorize, analyze and summarize the subjective and sentiment content of digital libraries.

This paper introduces a new general-purpose sentiment lexicon called the Wee Kim Wee School of Communication & Information (WKWSCI) Sentiment Lexicon, and reports an evaluation of its effectiveness in document-level and sentence-level sentiment categorization of a product-review corpus. The sentiment lexicon is not derived from a particular corpus, and is not specific to a particular domain. It is based on the *12dicts* common American English word lists compiled by Alan Beale from twelve source dictionaries—eight English-as-a second-language dictionaries and four "desk dictionaries" [2]. Specifically, we make use of Beale's *6of12* list comprising 32,153 American English words common to 6 of the 12 source dictionaries. This

R.B. Allen et al. (Eds.): ICADL 2015, LNCS 9469, pp. 82–93, 2015.
DOI: 10.1007/978-3-319-27974-9_9

reflects the core of American English vocabulary. Currently, the WKWSCI Sentiment Lexicon comprises adjectives, adverbs and verbs. Sentiment coding of nouns is in progress.

The project started three years ago when, dissatisfied with the sentiment lexicons that were available on the Web, we decided to systematically develop our own general-purpose sentiment lexicon. Since then, however, other researchers have developed their own sentiment lexicons, and a new version of SentiWordNet has been published.

This paper compares the WKWSCI lexicon with three comparable sentiment lexicons available on the Web:

1. General Inquirer[1]
2. MPQA (Multi-perspective Question Answering) lexicon[2]
3. Hu & Liu Lexicon[3]

We compare the effectiveness of the four lexicons in an automatic sentiment categorization task, using an Amazon product reviews dataset. In the experiments, we applied each lexicon to predict the document-level sentiment polarity (i.e. the overall rating assigned by the reviewer, converted to binary values of positive/negative), as well as sentence-level sentiment polarity.

There are two main approaches to automatic sentiment categorization: the machine learning approach and the lexicon-based approach.

A machine learning approach builds a sentiment categorization model using a training corpus. This approach basically selects words (or assigns weights to words) that are useful in distinguishing between positive and negative documents, based on a set of training documents that have been annotated with the sentiment category to predict (i.e. positive or negative). Usually, individual words in the documents are used as features in the model, and hence it is referred to as a bag-of-words approach. The most commonly-used machine learning methods in sentiment analysis are the Support Vector Machine (SVM) [3,4] and the Naïve Bayes method [5]. Wang and Manning [6] found the Naïve Bayes method to be more effective for snippets or short reviews, whereas SVM was more effective for longer documents or full-length reviews.

A sentiment lexicon-based approach uses a general-purpose or domain-specific sentiment dictionary, comprising a list of words, each word tagged as positive, negative or neutral (and sometimes with a value reflecting the sentiment strength or intensity). The lexicon may be developed manually [7,8], automatically using word associations with known "seed words" in a corpus [9,10], or semi-automatically deriving sentiment values from resources such as WordNet [11,12]. To predict the overall sentiment of a document, a formula or algorithm is needed to aggregate the sentiment values of individual words in the document to generate the document-level sentiment score.

Sentiment categorization models developed using machine learning are expected to be more accurate than a general-purpose sentiment lexicon, if the training corpus is

[1] http://www.wjh.harvard.edu/~inquirer/spreadsheet_guide.htm
[2] http://mpqa.cs.pitt.edu/lexicons/subj_lexicon/
[3] http://www.cs.uic.edu/~liub/FBS/sentimentanalysis.html#lexicon

sufficiently large. A model developed using machine learning is customized to the vocabulary of the corpus and the writing style of the genre. The machine learning approach has the disadvantage that a sufficiently large training corpus that has been annotated with the target sentiment category must be available or has to be constructed. With the proliferation of product review sites with user comments and ratings, there is an abundance of such annotated documents on the Internet. A machine learning approach is not feasible when there is no readily available annotated corpus.

A machine learning approach is also more appropriate for document-level sentiment categorization, where there are more textual features (i.e. words) to make sentiment category predictions. To perform finer sentiment analysis at the sentence or clause level, a sentiment lexicon is needed. Fine-grained sentiment analysis includes aspect-based sentiment analysis (identifying the writer's sentiment towards various aspects of a product or topic, rather than the overall sentiment) [13], multi-perspective sentiment analysis (identifying the sentiment of various stakeholders or roles) [14], and identifying the type of sentiment (rather than just positive or negative sentiment polarity) [15].

The disadvantage of lexicon-based methods is that words can have multiple meanings and senses, and the meaning and sense that is common in one domain may not be common in another. Furthermore, words that are not generally considered sentiment-bearing can imply sentiments in specific contexts. However, when a domain specific lexicon is not available, a good general-purpose sentiment lexicon will be useful, and can give acceptable results. Taboada et al. [8] developed a Semantic Orientation Calculator for computing the sentiment polarity and strength of words and phrases based on a manually-built sentiment lexicon, and showed that such a method is robust and can give reasonably good results across domains.

2 WKWSCI Sentiment Lexicon: Overall Characteristics

The WKWSCI Sentiment Lexicon was manually coded by 12 undergraduate students in the Bachelor of Communication program at the Wee Kim Wee School of Communication & Information, Nanyang Technological University, Singapore. Second-year and third-year undergraduate students were recruited to do the coding in the summer of 2013 and 2014. Students who responded to an email recruitment advertisement were given a coding test, and in each year, six students with the highest scores in the test were recruited. Each word list was coded by three coders. The sentiment coding was carried out in two phases:

- Phase 1: the coders coded the words as positive, neutral or negative. They were instructed to follow their first impression without agonizing over their coding, and to select "neutral" when in doubt. The codings that were not unanimous among the three coders were reviewed by the first author, who made the final decision. The implication of this approach is that some slightly positive and slightly negative words are coded as neutral.

- Phase 2: the words that were coded as positive in Phase 1 were subjected to a second-round coding by 3 coders into 3 subcategories: slightly positive (sentiment value of 1), positive (2) and very positive (3). Another 3 coders coded the negative words into slightly negative (-1), negative (-2) and very negative (-3). Again, the words that did not obtain unanimous sentiment values from the 3 coders were reviewed by the first author.

Nearly 16,000 words, comprising approximately 7,500 adjectives, 2,500 adverbs and 6,000 verbs, have been coded with sentiment values. Table 1 lists the number of adjectives, adverbs and verbs coded with the different sentiment values. There are 2,334 positive words, 4,384 negative words and 9,206 neutral words. It is noted that there are almost twice as many negative words as positive words in the lexicon. In contrast, there are usually more instances of positive words than negative words in a text corpus. Few words in the lexicon are very positive or very negative.

Looking at the distribution of verbs: there are many more negative verbs (1,284) than positive verbs (269). Furthermore, sentiment verbs tend to have weak sentiment strength: there are more than twice as many slightly negative verbs than negative and very negative verbs, and three times as many slightly positive verbs than positive and very positive verbs.

Some words have multiple parts-of-speech:

- 474 words occur as both adjectives and verbs
- 374 words occur as both adjectives and adverbs
- 83 words occur as both adverbs and verbs
- 65 words occur as adjectives, adverbs and verbs.

There are 177 words with multiple parts-of-speech that have conflicts in their sentiment score for the different parts-of-speech. Most of the conflicts involve a positive or negative sentiment for one part-of-speech, and neutral sentiment for another part-of-speech. There are two exceptions: “keen” and “smart” have positive sentiment as adjectives, but negative sentiment as verbs.

Table 1. Frequency of words in the WKWSCI lexicon, with various parts-of-speech and sentiment values

Sentiment Polarity	**Positive**			**Negative**			**Neutral**	**Total**
Sentiment Score	**3**	**2**	**1**	**-3**	**-2**	**-1**	**0**	
Adjective	60	686	735	34	1031	1318	3656	7520
Adverb	5	316	243	12	429	276	1091	2392
Verb	4	63	202	12	400	872	4459	6012
Total	2334			4384			9206	15924

3 Comparison with Other Sentiment Lexicons

We compared our lexicon with three other comparable sentiment lexicons available on the Internet. We excluded SentiWordNet[4] from the study as we did not find the lexicon effective enough in identifying sentiment polarity in an earlier project [13]. This was possibly because the use of SentiWordNet requires effective word sense disambiguation, which we did not attempt. However, a new version of SentiWordNet 3.0 has been published, which we shall evaluate in the future.

The General Inquirer [7] has 11,789 word senses (some words have multiple senses), grouped into 182 categories. In this study, we analyzed only those words in the categories Postiv (1915 words) and Negativ (2291). Furthermore, we compared only the words tagged Modif (which are mainly adjectives, with a few adverbs and verbs) and SUPV (which are mostly verbs). We analyzed the conflicts in sentiment coding between General Inquirer and WKWSCI Lexicon. The main conflicts are between neutral words in the WKWSCI lexicon which are coded as positive or negative in General Inquirer.

The MPQA Subjectivity Lexicon has 8,222 words: 2719 positive, 4914 negative and 591 neutral words. It includes adjectives, adverbs, verbs, nouns and "anypos" (any part-of-speech). This study compares the adjectives, adverbs and verbs with our WKWSCI lexicon, ignoring the nouns. The lexicon was aggregated from a variety of sources, including manually developed and automatically constructed sources. A majority of the entries were collected in a project reported by Riloff and Wiebe [16].

As with the General Inquirer, most of the conflicts are between neutral codings in WKWSCI and positive/negative codings in MPQA. There are 45 words that appear in both lexicons but with opposite polarity (i.e. ignoring neutral codings in WKWSCI lexicon). 22 positive words in MPQA are coded negative in WKWSCI, and 23 negative words coded positive in WKWSCI. The main reason for the conflicting polarities is multiple senses of words: a word can have a positive sense and a negative sense. Examples are *gritty, accountable, comical, dogged, edgy, eternal, expedient, formidable, imposing, rigorous, sharp, sober, sympathetic, uneventful, unobserved,* and *zealous*. These are coded "1" (slightly positive) in WKWSCI lexicon and negative in MPQA.

The sentiment coding in some cases depends on the narrow or broader context being considered. For example, to "commiserate" and to "empathize" are polite gestures (positive) in a narrow context, but they indicate a broader context of misfortune for the person being commiserated or empathized with. The coding in WKWSCI is biased towards the narrow context. Of the 173 neutral words (of any part-of-speech) in WKWSCI that match with words in MPQA, 47 are coded positive in MPQA and 47 coded negative. The MPQA coding looks reasonable. As the coders for the WKWSCI lexicon had been instructed to code a word as neutral when in doubt, they were quite conservative in assigning sentiment polarity.

The Hu & Liu lexicon [12] has 6,790 words with no part-of-speech tags: 2006 positive words and 4783 negative words. This lexicon was generated automatically using

[4] http://sentiwordnet.isti.cnr.it/

machine learning techniques based on customer reviews from various domains compiled over several years. Again, most of the conflicts with the WKWSCI lexicon involve neutral words in WKWSCI coded as positive or negative in the Hu & Liu Lexicon.

4 Evaluation Experiments and Results

4.1 Evaluation Corpus

The sentiment categorization experiments made use of a subset of an Amazon product review corpus, downloaded from `http://www.cs.uic.edu/~liub/FBS/sentimentanalysis.html`. The corpus was constructed by Jindal and Liu [17] for their study of opinion spam (fake review) detection. They noted that the corpus can be used for sentiment analysis experiments. The dataset has 25 product categories, each with up to 1000 positive and 1000 negative reviews. Each review is labelled as positive if the user rating score is 4 or 5, and negative if the user rating score is 1 or 2. We randomly selected 5 product categories out of 10 categories that have 1000 positive and 1000 negative reviews. The selected product categories are: apparel, electronics, kitchen & housewares, sports and outdoors, and video. For developing and evaluating machine learning models, we randomly selected 700 positive and 700 negative reviews from each product category to form the training set, and used the rest as the test set. This evaluation study made use of the review texts and the sentiment polarity (positive/negative). The review texts were lemmatized and tagged with part-of-speech tags using the Stanford core NLP parser [18].

We carried out the evaluation both at the document level and sentence level. For the sentence level evaluation, we randomly selected 50 positive and 50 negative reviews for each topic (500 reviews in all), and hired undergraduate students to code the sentences. Natural language toolkit 3.0 sentence tokenizer [18,19] was used to segment the review texts into sentences. Each sentence was coded by two coders, and conflicts were reviewed by the first author, who made the final decision.

It is important to note that three of the lexicons in the study were not constructed specifically for analyzing product reviews, whereas the Hu & Liu lexicon was developed based on product review texts. We were particularly interested to find out whether general-purpose sentiment lexicons can be applied with reasonable results to another domain.

4.2 Evaluation of Document-Level Sentiment Categorization

Two baseline experiments were carried out:

- Method 1a: Machine learning using bag-of-words, using Support Vector Machine (SVM) and Naïve Bayes method
- Method 1b: Lexicon-based method using *number of positive words – number of negative words.*

The SVM and Naive Bayes packages in the Scikit Learn Library [20] were used to develop the sentiment classifiers. The default parameters were used for both packages, with the SVM kernel set to polynomial.

Method 1a: Baseline machine learning method using bag-of-words. SVM and Naïve Bayes classifiers were built using the training dataset. Two weighting schemes were used: term frequency (tf) and term frequency*inverse document frequency (tf*idf). The results are given in Table 2. It can be seen that the results are about the same for tf and tf*idf weighting schemes, and for SVM and Naïve Bayes models. The accuracy rate is generally 81%; the highest accuracy is 82% for Naïve Bayes model using tf*idf weighting.

Table 2. Evaluation of Baseline SVM and Naïve Bayes models

	SVM (*tf* weighting)		**SVM (*tf**idf* weighting)**	
	Positive	**Negative**	**Positive**	**Negative**
Precision	0.807	0.808	0.789	0.836
Recall	0.809	0.807	0.848	0.773
F1 Score	0.808	0.807	0.818	0.803
Accuracy	0.808		0.811	
	Naïve Bayes (*tf*)		**Naïve Bayes (*tf**idf*)**	
	Positive	**Negative**	**Positive**	**Negative**
Precision	0.825	0.794	0.875	0.782
Recall	0.784	0.834	0.751	0.892
F1 Score	0.804	0.814	0.808	0.833
Accuracy	0.810		0.822	

Method 1b: Baseline lexicon-based method. The lexicon-based baseline method calculates sentiment scores for the reviews using the simple formula: *number of positive words - number of negative words.* The reviews are then ranked in decreasing score, and the top half of the reviews are categorized as positive, and the bottom half negative.

The results are given in Table 3. It can be seen that the WKWSCI lexicon (excluding slightly positive and slightly negative words) performed slightly better than the Hu & Liu lexicon, and clearly better than MPQA and General Inquirer. The WKWSCI lexicon obtained an accuracy rate of 72%, even though the lexicon is not derived from product review texts and does not include nouns.

Table 3. Accuracy of document-level sentiment categorization using baseline scoring method of counting positive and negative words

Lexicon	**Accuracy**
WKWSCI	0.694
WKWSCI (excluding slightly positive and slightly negative words)	0.723
Hu & Liu lexicon	0.710
MPQA	0.682
General Inquirer	0.634

Method 2: Lexicon-based method using logistic regression to determine the weights for different categories of words. Instead of just counting the number of positive and negative words, this method assigns different weights to different categories of words: each category is a combination of part-of-speech and sentiment strength (i.e. very positive, positive, etc.). Each word category then represents a feature whose value is the number of words of that category found in the document, normalized by dividing by the length of the review (i.e. review word count). Logistic regression (in the SPSS statistical package) is applied to the training dataset to determine the appropriate weights for each word category.

The logistic regression model for the WKWSCI lexicon indicates that the baseline score (using the baseline formula of Method 1b) should be combined with a normalized version of the baseline score (by dividing by the length of the review). The model also suggests that a higher number of adverbs indicates a negative review. For the WKWSCI lexicon, the accuracy improved from 0.723 for the baseline model (Model 1b) to 0.755 (see Table 4).

The logistic regression model for the Hu & Liu lexicon (not listed due to space constraints) indicates that the normalized version of the baseline score gives better results than the baseline score. In addition, the number of positive words have a significant impact on the accuracy of the sentiment categorization. The accuracy of the logistic regression model improved from 0.71 for the baseline model to 0.733 (see Table 4), which is still a little worse than the WKWSCI lexicon.

However, both models are substantially worse than bag-of-words machine learning models, which easily obtained accuracies of above 80%. The accuracy of the best machine-learning model probably represents the upper bound of what can be achieved using sentiment lexicons. The strength of sentiment lexicons is that training is not absolutely necessary, as the baseline scoring method still gives reasonable results of above 70%.

Table 4. Results of the logistic regression models for WKWSCI lexicon compared with the Hu & Liu lexicon on the test set

	WKWSCI lexicon	
	Positive reviews	**Negative reviews**
Precision	0.771	0.739
Recall	0.742	0.768
F1 Score	0.756	0.753
Accuracy	0.755	
	Hu & Liu lexicon	
	Positive reviews	**Negative reviews**
Precision	0.744	0.723
Recall	0.711	0.755
F1 Score	0.727	0.739
Accuracy	0.733	

4.3 Evaluation of Sentence-Level Sentiment Categorization

50 positive and 50 negative reviews were randomly sampled from each of the five topics, to make up 500 reviews in all. 1840 sentences were extracted from the 250 positive reviews, and 1528 sentences were extracted from the 250 negative reviews. They were coded by two coders into positive, negative and neutral/indeterminate sentiment polarity. Only unanimous codings were accepted as positive and negative sentences. There were 869 clearly positive sentences, and 964 clearly negative sentences. 24 reviews did not have any positive or negative sentences, and were dropped from the evaluation dataset.

To find out how important sentence-level sentiment is in determining the overall sentiment of a review, we calculated a sentiment score for each review using the formula: *number of positive sentences – number of negative sentences.* The reviews with a score of 0 and above were categorized as positive, and reviews with a score of -1 and below were categorized as negative. This obtained an accuracy rate of 0.937—for predicting the overall sentiment polarity of a review based on the number of positive and negative sentences. This indicates that accurate sentence-level sentiment categorization can improve the accuracy of document-level sentiment categorization.

Method 3: Baseline lexicon-based method for sentence categorization. The lexicon-based baseline method calculates sentiment scores for sentences using the simple formula: *number of positive words - number of negative words.* The accuracy of the sentence-level sentiment categorization is summarized in Table 5. Hu & Liu lexicon had the highest accuracy of 0.732, with WKWSCI obtaining the second highest accuracy of 0.716.

Table 5. Accuracy of sentence-level sentiment categorization using baseline scoring method

Lexicon	Accuracy
WKWSCI	0.716
WKWSCI (excluding slightly positive and slightly negative words)	0.692
Hu & Liu lexicon	0.732
MPQA	0.702
General Inquirer	0.669

Method 4: Lexicon-based method but using logistic regression to determine the weights for different categories of words. Stepwise logistic regression was applied to the training dataset to determine the appropriate weights for the number of positive words in the sentence, number of negative words, number of negation words, and the interaction variables—number of negation words multiplied by each of the other variables. The results of applying the logistic regression models for the four lexicons to the test dataset are given in Table 6. The accuracy for the Hu & Liu lexicon improved from 0.732 for the baseline model to 0.774 for the logistic regression model. The accuracy for WKWSCI lexicon improved from 0.716 to 0.758, which is a little worse than the results for Hu & Liu lexicon.

Table 6. Results for sentence-level sentiment categorization using logistic regression models for the four sentiment lexicons

	Polarity	Precision	Recall	F1-Score	Accuracy
WKWSCI	Positive	.772	.695	.732	**.758**
	Negative	.748	.815	.780	
MPQA	Positive	.732	.707	.719	.738
	Negative	.744	.767	.755	
General Inquirer	Positive	.700	.737	.718	.726
	Negative	.751	.715	.733	
Hu & Liu	Positive	.838	.650	.732	**.774**
	Negative	.737	.886	.805	

4.4 Error Analysis

We carried out an error analysis of the false positive and false negative errors that had been predicted with high probability of above 0.70 by the logistic regression model. 27 sentences were incorrectly predicted by the model to be negative with high probability, and 22 sentences were incorrectly predicted to be positive with high probability.

The biggest source of error is the need for common sense inferencing to identify a review as positive or negative. This is especially true in the case of false positives. Several of the cases involve long, complex sentences in reviews of videos. Users also tend to use sarcasm or hyperbole to express negative sentiments. Inferencing is difficult to model using lexicon-based methods or bag-of-words machine-learning models.

The second major source of the error for false negatives is the incorrect handling of negation words. Our regression models do take into consideration the presence of negation words in the sentence, but they are handled as an independent negation feature and as interactions with the sentiment features. In other words, we did not consider the position of the negation word—whether it immediately precedes a sentiment-bearing word. From our observation, the negation word usually precedes the sentiment-bearing word that it modifies, but there can be up to 2 words in between. Surprisingly, negation handling did not appear to be a major problem in false positive predictions.

Sentiment-bearing phrases is the third source of error. They include: *be careful, do not bother, just like any other, hard to go wrong*, and *cannot beat it*. This can be addressed by reviewing 2 or 3-word sequences associated with positive or negative reviews, and compiling a list of such sentiment phrases.

We also examined the 74 sentences that do not have any word matches with the four lexicons. The majority of the cases (37) require commonsense inferencing, though some of these can be handled using domain-specific cue phrases that indicate negative features or sentiments. 15 of the cases contained the words “do not buy”, “never ever buy”, “never buy”, or “not buy”. 5 cases involve colloquial sentiment expressions such as “dorky”, “wtf”, “this movie rocks”, “what a crock” and “yikes”.

5 Conclusion

We have described the characteristics of the WKWSCI sentiment lexicon in comparison with three other comparable lexicons. The WKWSCI lexicon currently covers adjectives, adverbs and verbs from the *6of12dict* lexicon. The sentiment coding was carried out by undergraduate students in the Bachelor of Communication program. Each word was reviewed by 3 coders, and assigned a sentiment strength on a 7-point scale. Sentiment-bearing nouns are currently being coded.

From direct comparisons between the WKWSCI lexicon and the other lexicons, it was found that the WKWSCI lexicon is weaker in the category of slightly positive and slightly negative words, as the coders were instructed to assign the neutral category in cases of doubt. We thus recommend that the WKWSCI lexicon be supplemented with a list of words with the same sentiment polarity in both MPQA and the Hu & Liu Lexicon.

The four lexicons were used to perform document-level sentiment categorization on an Amazon product reviews dataset, as well as sentence-level sentiment categorization of sentences from a subset of the reviews. For document-level sentiment categorization, the WKWSCI lexicon performed slightly better than the Hu & Liu lexicon, whereas Hu & Liu performed slightly better for sentence-level sentiment categorization. The WKWSCI lexicon obtained an accuracy of 72% using a simple count of positive and negative words. The accuracy increased to 75% when the weights for the various counts were determined using logistic regression. For sentence-level sentiment categorization, the WKWSCI lexicon also obtained an accuracy of 75% when the weights were determined using logistic regression. However, the Hu & Liu lexicon did better this time, obtaining 77% accuracy.

Both WKWSCI lexicon and the Hu & Liu lexicon are clearly better than MPQA and General Inquirer. The Hu & Liu lexicon was derived from product review texts and is thus customized for the domain. It also does not have part-of-speech tags and thus includes nouns in the lexicon. In contrast, the WKWSCI lexicon is general-purpose and currently does not include nouns. Sentiment coding of nouns is in progress.

Bag-of-words machine-learning categorization models were developed using Support Vector Machine and Naïve Bayes machine learning methods. These models obtained an accuracy of about 82% for document-level sentiment categorization. This probably represents the upper bound of what can be achieved using a lexicon-based method. The strength of the lexicon-based method is in sentence-level and aspect-based sentiment analysis, where it is difficult to apply machine-learning because of the small number of features. For document-level sentiment categorization, sentiment lexicons can obtain reasonable results in different domains using simple counts of positive and negative words, without training. However, more work is needed to confirm this across a variety of domains and text genres.

References

1. Oomen, J., Aroyo, L.: Crowdsourcing in the cultural heritage domain: opportunities and challenges. In: Proceedings of the 5th International Conference on Communities and Technologies, pp. 138–149. ACM, June 2011
2. Dicts introduction. http://wordlist.aspell.net/12dicts-readme/

3. Cortes, C., Vapnik, V.: Support-Vector Networks. Machine Learning **20**(3), 273–297 (1995)
4. Vapnik, V.N.: Statistical Learning Theory. John Wiley and Sons, New York (1998)
5. Zhang, H.: The optimality of Naive Bayes. In: Proceedings of the Seventeenth Florida Artificial Intelligence Research Society Conference, pp. 562–567. The AAAI Press (2004)
6. Wang, S., Manning, C.D.: Baselines and bigrams: simple, good sentiment and topic classification. In: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, pp. 90–94. Association for Computational Linguistics (2012)
7. Stone, P.J., Dunphy, D.C., Smith, M.S., Ogilvie, D.M.: The General Inquirer: A Computer Approach to Content Analysis. MIT Press, Cambridge (1966)
8. Taboada, M., Brooke, J., Tofiloski, M., Voll, K., Stede, M.: Lexicon-Based Methods for Sentiment Analysis. Computational Linguistics **37**(2), 267–307 (2011)
9. Hatzivassiloglou, V., McKeown, K.: Predicting the semantic orientation of adjectives. In: Proceedings of 35th Meeting of the Association for Computational Linguistics, pp. 174–181 (1997)
10. Turney, P., Littman, M.: Measuring Praise and Criticism: Inference of Semantic Orientation from Association. ACM Transactions on Information Systems **21**(4), 315–346 (2003)
11. Esuli, A., Sebastiani, F.: SentiWordNet: a publicly available lexical resource for opinion mining. In: Proceedings of 5th International Conference on Language Resources and Evaluation (LREC), pp. 417–422 (2006)
12. Hu, M., Liu, B.: Mining and summarizing customer reviews. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD–2004), pp. 168–177. ACM, New York (2004)
13. Thet, T.T., Na, J.C., Khoo, C.: Aspect-Based Sentiment Analysis of Movie Reviews on Discussion Boards. Journal of Information Science **36**(6), 823–848 (2010)
14. Wiebe, J., Wilson, T., Cardie, C.: Annotating Expressions of Opinions and Emotions in Language. Language Resources and Evaluation **39**(2–3), 165–210 (2005)
15. Khoo, C., Nourbakhsh, A., Na, J.C.: Sentiment Analysis of News Text: A Case Study of Appraisal Theory. Online Information Review **36**(6), 858–878 (2012)
16. Riloff, E., Wiebe, J.: Learning extraction patterns for subjective expressions. In: Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing (EMNLP–2003), pp. 105–112. Association for Computational Linguistics (2003)
17. Jindal, N., Liu, B.: Opinion spam and analysis. In: Proceedings of the 2008 International Conference on Web Search and Data Mining, pp. 219–230. ACM, New York (2008)
18. Manning, C.D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S.J., McClosky, D.: The stanford CoreNLP natural language processing toolkit. In: Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pp. 55–60 (2014)
19. Bird, S., Loper, E., Klein, E.: Natural Language Processing with Python. O'Reilly Media (2009)
20. Pedregosa, F., et al.: Scikit-learn: Machine learning in Python. The Journal of Machine Learning Research **12**, 2825–2830 (2011)