Article

Journal of Information Science
1–21

DOI: 10.1177/0165551510000000
jis.sagepub.com

Lexicon-Based Sentiment Analysis: Comparative Evaluation of Six Sentiment Lexicons

Christopher S.G. Khoo
Wee Kim Wee School of Communication & Information, Nanyang Technological University, Singapore

Sathik Basha Johnkhan
Wee Kim Wee School of Communication & Information, Nanyang Technological University, Singapore

Abstract
This paper introduces a new general-purpose sentiment lexicon called WKWSCI Sentiment Lexicon, and compares it with five existing lexicons: Hu & Liu Opinion Lexicon, MPQA Subjectivity Lexicon, General Inquirer, NRC Word-Sentiment Association Lexicon and SO-CAL lexicon. The effectiveness of the sentiment lexicons for sentiment categorization at the document-level and sentence-level was evaluated using an Amazon product review dataset and a news headlines dataset. WKWSCI, MPQA, Hu & Liu and SO-CAL lexicons are equally good for product review sentiment categorization, obtaining accuracy rates of 75% to 77% when appropriate weights are used for different categories of sentiment words. However, when a training corpus is not available, Hu & Liu obtained the best accuracy with a simple-minded approach of counting positive and negative words, for both document-level and sentence-level sentiment categorization. The WKWSCI lexicon obtained the best accuracy of 69% on the news headlines sentiment categorization task, and the sentiment strength values obtained a Pearson correlation of 0.57 with human assigned sentiment values. It is recommended that the Hu & Liu lexicon be used for product review texts, and the WKWSCI lexicon for non-review texts.

Keywords
Sentiment lexicon; Sentiment analysis; Sentiment categorization

1. Introduction

Automated sentiment analysis is attracting substantial interest from public and corporate organizations seeking to mine user-reviews and social media content for customer sentiment and opinion towards their products and services. Researchers are actively developing techniques for various kinds of sentiment analysis. A basic kind of sentiment analysis is sentiment categorization—categorizing pieces of text into positive and negative sentiment polarity (or valence or orientation). Researchers have investigated sentiment categorization at the document level (including product reviews), as well as sentence level and even text passage level (including phrases and clauses).

There are two basic approaches to automatic sentiment categorization—machine-learning approach and lexicon-based approach. Machine-learning methods often employ a "bag-of-words" approach of using words (usually lemmatized or stemmed) in the corpus as independent features in a feature vector to represent the documents. Multiword terms are also sometimes used as features. The value for each word or term feature is usually taken as the term frequency *tf* (i.e. number of occurrences of the word in the document), or *tf***idf* (*tf* multiplied by the inverse document frequency—the number of documents containing the term). Supervised learning is used to develop a classifier to classify a document into positive or negative sentiment, sometimes with a confidence value that can be used as an indication of sentiment strength. In effect, supervised machine learning is used to identify word features that are effective in distinguishing between positive and negative sentiments. The most commonly-used machine learning methods in sentiment analysis are the Support Vector Machine (SVM) [1,2] and the Naïve Bayes method [3]. Wang and Manning [4] 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.

Corresponding author:
Christopher Khoo, Wee Kim Wee School of Communication & Information, Nanyang Technological University, 31 Nanyang Link, Singapore 637718
Email: chriskhoo@pmail.ntu.edu.sg

The second approach, the focus of this study, makes use of an existing lexicon with words or multiword terms tagged as positive, negative or neutral (sometimes with a value reflecting the sentiment strength or intensity). The lexicon may be developed manually, for example [5,6], automatically using word associations with known "seed words" in a corpus, or example [7,8], or semi-automatically deriving sentiment values from resources such as WordNet, for example [9]. 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.

The advantage of the machine-learning approach is that the role of each word in the sentiment categorization process is customized to the corpus and application. It is well-known that the meaning of words and the sentiment they reflect depend to some extent on the domain and context [10,11]. The disadvantage of the machine-learning approach is that a sizeable training corpus is needed to develop the classifier. With the proliferation of product review sites with user comments and ratings, there is an abundance of such annotated documents on the Internet. When a sizeable training corpus is not available, an existing sentiment lexicon is needed for sentiment categorization.

A machine learning approach is also more appropriate for document-level sentiment categorization, where there are more textual features (i.e. words) on which to base the 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) [12], multi-perspective sentiment analysis (identifying the sentiment of various stakeholders or roles) [13], and identifying the type of sentiment (rather than just positive or negative sentiment polarity) [14]. Researchers (e.g., Taboada et al. [6]) have shown that general-purpose sentiment lexicons give reasonably good results when applied to different corpora, and can also be used as seed sets to derive domain-specific sentiment-bearing words from a corpus.

The disadvantage of using a lexicon 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. Given a sufficiently large training corpus, a machine learning model is expected to outperform a lexicon-based model.

This study focused on the lexicon-based approach to sentiment categorization of text at the document and sentence level. In particular, it compared the effectiveness of six sentiment lexicons in performing sentiment categorization on a product review corpus at the document and sentence level, as well as on a news headlines corpus. The six sentiment lexicons investigated were:

(1) General Inquirer[1]
(2) MPQA (Multi-perspective Question Answering) Subjectivity Lexicon[2]
(3) Hu & Liu Opinion Lexicon[3]
(4) National Research Council Canada (NRC) Word-Sentiment Association Lexicon[4]
(5) Semantic Orientation Calculator (SO-CAL) lexicon version 1.11[5]
(6) WKWSCI Sentiment Lexicon that we developed.

We dropped SentiWordNet 3.0[7] from the study as it performed poorly in the baseline evaluation, described later. We compared the effectiveness of the lexicons against one another, as well as in comparison to bag-of-words models using Support Vector Machine and Naïve Bayes machine learning methods. In the process, we investigated whether lexicons developed for one domain can be effective in another domain.

It is well-known that people's sentiment can have different intensities or strengths, often represented as star ratings on product review sites. Automated sentiment analysis can assign both a sentiment polarity and sentiment strength value to a text passage. Two of the lexicons we evaluated (SO-CAL and WKWSCI) contain sentiment strength values for each word. We applied these two lexicons to a news headlines corpus (SemEval-2007 Task 14 "Affective Text" [15]) that had been annotated with sentiment strength values. This is to assess the effectiveness of the two lexicons in determining the sentiment strength of the text.

The motivation for this comparative study is our development of a new general-purpose sentiment lexicon called *WKWSCI Sentiment Lexicon* (named after our school). The lexicon is based on the *12dicts* common American English word lists compiled by Alan Beale from twelve source dictionaries [16]. Specifically, we made use of Beale's *6of12* list comprising 32,153 American English words common to 6 of the 12 source dictionaries. This reflects the core of American English vocabulary. We embarked on the project in the summer of 2012 when we were dissatisfied with the sentiment lexicons that were publicly available at the time. Our lexicon was completed at the end of 2016. Meanwhile, other researchers have developed their own sentiment lexicons, and a new version of SentiWordNet 3.0 was published. The

results of this study indicate that different sentiment lexicons, developed using different methods, have different characteristics and yield different results when applied to different types of texts.

An evaluation of an earlier incomplete version of the WKWSCI Sentiment Lexicon (without nouns) was reported in [17].

2. Related works

As manual construction of a sentiment lexicon is labor intensive and time consuming, much of the research literature on sentiment lexicon has focused on ways of constructing it and adapting it to different domains with less effort. There are four main approaches to sentiment lexicon construction:

(1) Manual construction, including crowdsourcing and gamification
(2) Bootstrapping from a set of seed words, using lexical, syntactic or associative relations
(3) Adapting a lexicon from another domain through some kind of transfer learning
(4) Machine learning or probabilistic learning, based on human sentiment coding or star rating of bigger chunks of text (e.g., sentences, reviews or social media posts).

2.1. Manual construction of sentiment lexicon

To speed up manual lexicon construction, researchers are exploring the use of crowdsourcing to engage a pool of coders on Internet platforms such as Amazon's Mechanical Turk[7]. Mohammad and Turney [18] coded nearly 9,000 words with associated emotion categories and sentiment valence scores using Mechanical Turk. They noted two main issues in using crowdsourcing for sentiment coding:

(1) Design of the sentiment coding tasks: the coding task should be simple enough to be done without training and specialized knowledge; the instructions should be clear, brief and simple; and the task should be interesting enough to attract coders
(2) Quality control: there should be a means of identifying coders who are not doing the task properly because of misinterpretation or dishonesty.

They found that the way the coding instructions were phrased affected the coding quality: asking the coder if a term was "associated" with a particular emotion was found to yield better intercoder agreement than asking whether a term "evokes" an emotion. They designed each task (referred to as a "Human Intelligence Task" or a "HIT" in Mechanical Turk) to include a few questions, one of which was a word choice question asking which of four different words was closest in meaning to the target word. This question indirectly conveyed the intended word sense to be used for coding, as well as provided a means of checking whether the coder understood the word.

Gamification has been used to make the task more fun and more accurate at the same time. This is often accomplished using collaborative (multiplayer) games where the coder is rewarded for entering the same coding as another player (whether synchronously or asynchronously). Hong et al. [19] designed a two-player game called "Tower of Babel" involving players assigning a sentiment polarity to words by manipulating falling blocks, each labelled with a word, into three stacks representing positive, neutral and negative sentiment categories. The players were rewarded for assigning the same sentiment category as the partner. They used the game to code Korean sentiment words, and found the game to be as effective as manual coding but faster, and the coding experience was perceived more positively than manual coding.

Thisone et al. [20] designed a two-player game with one player taking the role of "suggester" and the other "guesser." The suggester was given a review document and asked firstly to assess its sentiment polarity, and secondly to select a single word (or short sequence of words) that best reflected the polarity assigned to the document. The guesser was then given the selected word (or sequence of words), and asked to guess the polarity of the document based on the word. This game thus accomplishes two tasks—constructs a sentiment corpus with document-level sentiment coding, and derives a sentiment lexicon from it.

2.2. Bootstrapping a sentiment lexicon from a set of seed words

Bootstrapping a sentiment lexicon from a set of seed words can be accomplished by making use of semantic relations between the seed words and other words in a lexical resource (e.g., WordNet or a thesaurus), or associative or syntactic relations with other words in a corpus.

Researchers have made use of semantic (i.e. paradigmatic) relations between seed words and other words in WordNet to propagate sentiment values. Kamps et al. [21] build a network of adjectives based on the synonym relation specified in WordNet. The sentiment polarity of an adjective was determined based on the shortest path to positive and negative seed adjectives ("good" and "bad"). Esuli and Sebastiani [9] also built a relational network of word senses based on the word glosses (i.e., definitions) given in WordNet, to construct the well-known SentiWordNet lexicon.

Another approach is to make use of word associations in a corpus, based on proximity (or co-occurrence within a specific window size) with a seed word, or based on syntactic relations with the seed word. In a seminal study, Hatzivassiloglou and McKeown [7] used a seed set of 1336 most frequent adjectives in a newspaper corpus, manually coded them with sentiment polarity, and extracted other adjectives from the corpus that were linked to these seed adjectives with the conjunctions "and", "or", "but", "either-or" and "neither-nor". They found that adjectives linked by conjunctions tended to have the same sentiment polarity (with about 78% probability), except for "but" which tended to link adjectives with opposite polarity (69% probability).

Turney and Littman [8] made use of associative relations in a corpus, based on co-occurrence in the same text neighborbood with a seed set of seven positive and seven negative adjectives. They investigated two methods of calculating the association strength between two words:

(1) Pointwise mutual information measure [22] of co-occurrence in different window sizes
(2) Cosine similarity between the word vectors—each word represented by a vector derived from its frequency of occurrence in the set of documents (with the number of document-dimensions reduced using latent semantic analysis).

They found both methods of measuring associative relationship to be effective, but the mutual information measure is easier to implement as the component counts can be obtained using an information retrieval system that can perform word-proximity searching,

A network of word relationships can be used to propagate sentiment values from seed words through the network, with the sentiment activation value attenuating as it travels further away from the seed words. However, activation values from multiple seed words can be aggregated. Glavaš et al. [23] built a network of words based on associative relations in a corpus. A set of 15 positive, 15 negative and 15 neutral words were used as seed words, and their sentiment values were propagated through the network.

2.3. Adapting a lexicon from another domain through transfer learning

As it is well-known that a sentiment lexicon needs to be customized for a domain to obtain optimal results, many authors have explored methods to optimize a general-purpose sentiment lexicon for a particular domain, or to adapt or extend a sentiment lexicon constructed for one domain to another domain. Many of the studies made use of sophisticated machine learning techniques [e.g., 24, 25]. Some studies used a general-purpose sentiment lexicon to perform a "first cut" sentiment analysis on a new corpus, from which new words associated with positive or negative documents were shortlisted. For example, Bahrainian et al. [26] first applied an existing sentiment lexicon to a Twitter corpus to calculate the sentiment values of Twits. They then identified new words that were strongly associated with positive or negative Twits. The sentiment values for the new words were determined by calculating the number of positive Twits containing the word minus the number of negative Twits. Thelwall and Buckley [11] used a similar approach, but made use of human coding to assign sentiment scores to social media posts, which were then compared to the scores assigned using a general-purpose lexicon. For a new target word, a sentiment value was assigned that would reduce the average difference between human coding and automated coding for posts containing the word.

2.4. Using machine-learning to identify sentiment-bearing words

Some researchers have used machine learning and probabilistic learning models to identify sentiment-bearing words. Yates et al. [27] and Xu et al. [28] used Latent Dirichlet Allocation [29], a probabilistic modelling technique, to model

the relationships between words, aspects, sentiments and documents. In this approach, documents are modelled as mixtures of topics (or aspects of a topic) and sentiments, which generate the words in the text.

2.5. Usefulness of general-purpose sentiment lexicons

The two studies that are closest to our work are the studies of Taboada et al. [6] and Thelwall et al. [30], who showed that their general-purpose sentiment lexicons gave reasonably good results when applied to different corpora and domains, including short documents (e.g., product reviews, blog posts, and various kinds of social media posts) and sentences. The SentiStrength system of Thelwall et al. assigned separate scores (on a scale of 1 to 5) for both positive and negative sentiments, on the assumption that a piece of text can express both positive and negative sentiments. On a range of social media texts (except product reviews), their system obtained a moderate correlation of 0.56 with human-assigned positive sentiment scores, and 0.57 with negative sentiment scores.

Taboada et al. [6] and Thelwall et al. [30] used sophisticated methods of calculating the sentiment strength of a piece of text, with careful handling of negation, intensifiers, mitigators (sometimes called "diminishers" and "downtoners"), irrealis markers and other linguistic features. In our opinion, the heuristics used in adjusting the prior word sentiment scores in the presence negation, intensifiers, etc., merit further study. As Taboada et al. noted, different intensifiers/mitigators modify the prior sentiment values of words differently: "extraordinarily" is a stronger intensifier than "rather", and intensification may be different for a word that has a strong sentiment valence than for one that is less intense (e.g., "truly fantastic" versus "truly okay").

Our study focuses on comparing lexicons, choosing to employ simple methods of calculating sentiment valence values that can be easily implemented with simple programming. We also adopt a simple-minded way of handling negation that improved sentiment categorization accuracy significantly.

3. Sentiment lexicons evaluated

The section gives an overview of the origin, size and features of the sentiment lexicons that were evaluated in this study, in comparison with the WKWSCI lexicon.

3.1. WKWSCI Sentiment Lexicon

The WKWSCI Sentiment Lexicon was manually coded by undergraduate students in the Bachelor of Communication Studies program at the Wee Kim Wee School of Communication & Information, Nanyang Technological University, Singapore. Third and fourth-year undergraduate students were recruited to do the coding in the summers of 2012 to 2015. 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 three coders into three subcategories: slightly positive (sentiment value of 1), positive (2) and very positive (3). Another three 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 three coders were reviewed by the first author.

The lexicon contains 29,718 words, comprising 3,121 positive words 7,100 negative words and nearly 19,500 neutral words. Table 1 lists the number of adjectives, adverbs, verbs and nouns coded with the different sentiment values. There are more than 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 distribution of verbs: there are many more negative verbs than positive verbs. Furthermore, sentiment-bearing 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:

- 473 words occur as both adjective and verb
- 374 words occur as both adjective and adverb
- 83 words occur as both adverb and verb
- 65 words occur as adjective, adverb and verb.

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

Sentiment Polarity			Positive			Negative	Neutral	Total
Sentiment Value	3	2	1	-1	-2	-3	0	
Adjective	60	692	750	1400	1040	34	3736	7712
Adverb	26	326	250	299	446	13	1154	2514
Verb	4	65	236	970	411	12	4531	6229
Noun	23	455	234	1338	1127	10	10076	13263
Total			3121			7100	19497	29718

Table 2. General statistics for the six sentiment lexicons, and number of words in common with WKWSCI Lexicon.

LEXICON	Positive				Neutral				Negative			
	Adj	Adv	Verb	Noun	Adj	Adv	Verb	Noun	Adj	Adv	Verb	Noun
WKWSCI	1502	602	305	712	3736	1154	4531	10076	2474	758	1393	2475
General Inquirer	771 (549)*	-	406 (170)	681 (270)	-	-	-	-	800 (603)	-	702 (475)	761 (449)
MPQA**	1171 (688)	128 (90)	380 (186)	677 (277)	235 (74)	19 (8)	76 (63)	144 (58)	1839 (1031)	183 (110)	869 (650)	1346 (632)
SO-CAL	1250 (707)	448 (285)	351 (164)	544 (266)	1 (-)	0 (-)	0 (-)	0 (-)	1576 (967)	429 (249)	791 (542)	1005 (596)
NRC				2312 (982)				-				3324 (2213)
Hu & Liu				2006 (1278)				-				4783 (2958)
SentiWordNet (weighted avg)	5273 (1030)	2179 (444)	2073 (177)	9052 (426)	9354 (1758)	1808 (583)	7803 (2810)	97007 (7147)	6852 (1656)	494 (151)	2373 (686)	11740 (1175)
SentiWordNet (sense #1)	4805 (913)	2078 (416)	1357 (130)	7672 (331)	10401 (2163)	1964 (651)	8508 (3567)	99706 (7930)	6273 (1488)	439 (136)	1664 (494)	10420 (997)

* Number in parenthesis indicate the number of words in agreement with WKWSCI.

** MPQA also has an "anyPOS" category with 362 positive words, 109 neutral words and 676 negative words.

There is no overlap between nouns and the other parts-of-speech by design: words that occur as both noun and another part-of-speech were eliminated from the list of nouns, in order to reduce the number of nouns to code.

There are 147 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 three exceptions: "keen", "smart" and "humble" have positive sentiment as adjectives, but negative sentiment as verbs.

Table 2 provides basic statistics on the other five lexicons and SentiWordNet 3.0, in comparison with the WKWSCI lexicon. Two versions of SentiWordNet 3.0 are included in the table—a version using a weighted average of the sentiment scores for the different senses of each word, and a version using the sentiment score of the first sense of the word. In most cases, at least half the words agree with the polarity coding in the WKWSCI lexicon. The main exception is SentiWordNet that has large numbers of words in each category, with only a small proportion agreeing with WKWSCI coding.

A majority of the conflicts involve neutral words in the WKWSCI lexicon which are coded positive or negative in the other lexicon. The conflicting polarities are due to words having different sentiments for different senses and in different contexts. We hypothesize that these have more sentiment ambiguity, and may need to be customized for different domains or applications. It is interesting that "proud" is coded negative in WKWSCI lexicon but positive in all the other lexicons. The sentiment orientation of "proud" depends on the context: it is negative in "proud, arrogant man", but positive in "proud owner of" and "proud father of".

3.2. General Inquirer

General Inquirer [6] 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 words). Constructed in the late 1960's, Devitt and Ahmad [31] explained that it is composed of frequency word lists from the Harvard IV Dictionary [32] and the Lasswell Dictionary [33]. The categories were hand-tagged essentially by "shrewd guesswork with numerous minor revisions" [34]. The General Inquirer categories were developed for social-science content-analysis research, and were added over time by various researchers.

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. Looking at the stronger conflicts involving words in the WKWSCI lexicon coded *2, 3, -2,* and *-3,* the positive words in WKWSCI that are coded negative in General Inquirer are "incredible" and "rigor". The negative words in WKWSCI that are coded positive in General Inquirer are "proud", "willful", "apocalypse", "impunity", and "notoriety". Without considering the context, the sentiment values assigned in WKWSCI look reasonable. One can of course find contexts in which "incredible" is negative, and "proud" is positive.

3.3. MPQA Subjectivity Lexicon

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). The lexicon was aggregated from a variety of sources, including manually developed sources as well as automatically constructed sources. A majority of the entries were collected in a project by Riloff and Wiebe [35].

As with the General Inquirer, most of the conflicts are between neutral codings in WKWSCI and positive/negative codings in MPQA. For the words coded *2, 3, -2* and *-3* in the WKWSCI lexicon, the conflicting words coded positive in WKWSCI are "incomparable", "incomparably", "doggedly", "rhapsodize" and "overcome" (verb), and the conflicting words coded negative in WKWSCI are "ingratiating", "joyless", "giddy", "indulgent", "terrified", "terrifying", and "truculent".

The sentiment polarity of some words depends on the narrow or broader context being considered. For example, "commiserate" and "empathize" (both coded *1* in WKWSCI) is a polite gesture (positive) in a narrow context, but they indicate a broader context of misfortune for the person being commiserated/empathized with. So the coding in WKWSCI is biased towards the narrow context.

3.4. Hu & Liu Opinion Lexicon

The Hu & Liu Opinion Lexicon [36] has 6,790 words with no part-of-speech tags: 2006 positive words and 4783 negative words. This lexicon was generated automatically using 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.

Positive words in WKWSCI with sentiment values *2* or *3* that conflict with the coding in Hu & Liu are: "incomparably", "rhapsodize", "flair", "gritty", "incomparable", "tenderness", "uproarious", and "uproariously". Negative words in WKWSCI with sentiment values *-2* and *-3* that conflict with Hu & Liu are "proud" and "unequivocal". We concede that "unequivocal" is incorrectly coded in the WKWSCI lexicon.

3.5. NRC Word-Sentiment Association Lexicon

The entries in the NRC lexicon were taken from three sources:

(1) Most frequent 200 unigrams and 200 bigrams for each part-of-speech (adjectives, adverbs, nouns and verbs) from the Macquarie Thesaurus [37], with the frequent terms identified by matching with Google n-gram corpus [38]
(2) 640 terms taken from the WordNet Affect Lexicon [39]
(3) Over 8,000 terms from General Inquirer.

Each term is coded with its association with eight emotion categories (anger, fear, anticipation, trust, surprise, sadness, joy, and disgust) and two sentiment categories (negative and positive). The coding was crowdsourced using Amazon's Mechanical Turk, and each term had valid codings from at least three persons. This study made use of the 2312 positive words and 3324 negative words in the lexicon.

Positive words in WKWSCI with sentiment values *2* or *3* that conflict with NRC coding are "conviction", "pomp", "rigor" and "touched". Negative words in WKWSCI with sentiment values *-2* or *-3* that conflict with NRC are "complacency" and "proud".

3.6. SO-CAL Lexicon Version 1.11

SO-CAL lexicon version 1.11 is a manually constructed lexicon that was coded by a native speaker, and reviewed by a committee of three researchers [6]. The main sources of the terms were:

(1) Adjectives taken from a corpus of 400 Epinions reviews, covering eight categories: books, cars, computers, cookware, hotels, movies, music, and phones
(2) A subset of 100 movie reviews from the Polarity Dataset [40]
(3) Positive and negative words from General Inquirer.

Each term is assigned a sentiment valence value on a scale of -5 to 5. A companion Semantic-Orientation Calculator (SO-CAL) program makes use of the lexicon to calculate a semantic valence score for text documents, taking into consideration negation, intensifiers, irrealis moods, etc. This study did not make use of the SO-CAL program, but only the lexicon. The 563 multiword terms and 169 regular expression patterns in the lexicon were also not included in this study, which focused on single words.

Positive words in WKWSCI with score *2* and *3* that conflicts with SO-CAL coding are "brave", "gritty", "ingenuous", "trendy", "fantastically", "thankfully", "glorify" and "pomp". Negative words in WKWSCI with score *-2* and *-3* that conflicts with SO-CAL are "hysterical", "ingratiating", "proud", "expensively" and "notoriety".

4. Evaluation experiments and results

4.1. Evaluation corpora

The sentiment categorization experiments made use of a subset of an Amazon product review corpus[8]. The corpus was constructed by Jindal and Liu [41] for their study of opinion spam (fake review) detection. They noted that the corpus can be used for sentiment analysis experiments as well. 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: apparels, 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 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 [42].

We carried out the evaluation of the sentiment categorization 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 [42,43] was

used to segment the review texts into sentences. Each sentence was coded by two coders, and only unanimous codings were accepted as positive and negative sentences.

The Amazon product review corpus carries only sentiment polarity coding of *positive/negative*. To evaluate sentiment scoring that includes sentiment strength values, we made use of a second corpus—the news headlines corpus used in SemEval-2007 Task 14 "Affective Text" [15]. The corpus comprises 1000 news headlines sampled from news websites and major newspapers (including New York Times, CNN, BBC News and Google News). The headlines were assigned scores on six emotion categories (Anger, Disgust, Fear, Joy, Sadness, Surprise). We made use only of the emotion valence coding that has values from *-100* (indicating highly negative) to *100* (highly positive). A value of 0 indicates neutral valence. Of the 1000 headlines, 468 have positive valence, 526 have negative valence, and 6 have neutral valence. The news headlines corpus allows us to test the effectiveness of the lexicons when applied to a different domain from product reviews.

It is important to note that five of the lexicons in the study were not constructed specifically for analyzing product reviews, whereas the Hu & Liu Lexicon were developed based on product review texts. As mentioned earlier, we were interested to find out whether sentiment lexicons built for a domain and application can be applied with reasonable results to another domain/application.

4.2. Negation handling

In an earlier study [44], we had used a dictionary of negation expression patterns (comprising sequences of part-of-speech tags) to identify negation expressions in the text. This was found to yield a small improvement in the accuracy of sentiment categorization from 76% to 79%, using a model developed by applying Support Vector Machine to a unigram bag-of-words representation. In this study, we decided to adopt a simpler method of handling negation. If a negation word (e.g., "not") occurs to the left of a sentiment-bearing word, with up to one word in between, the sentiment polarity is reversed and a weak sentiment strength is assigned. So, "not good" and "not so good" are considered slightly negative terms. In a preliminary experiment, we found that using a window size of 3 (i.e. with up to 1 word between the negation word and the sentiment-bearing word) is effective in identifying negated terms. A window size of 4 yields too many errors. The negation words used are: *barely, cease, hardly, neither, no, non, not, nothing, n't, prevent, rarely, seldom, stop* and *unlikely*.

4.3. Document-level sentiment categorization of product reviews

Two baseline experiments were carried out:

- Experiment 1a: Machine learning using bag-of-words, using Support Vector Machine and Naïve Bayes method
- Experiment 1b: Lexicon-based method using the following formula to calculate the sentiment score for a document: *number of positive words – number of negative words*, normalized by the number of words in the review.

One evaluation experiment was carried out:

- Experiment 2: Lexicon-based method but using logistic regression to determine the weights for different categories of words.

The results from the machine learning methods represent the upper-bound of what can be achieved using a general-purpose sentiment lexicon that has not been optimized for the corpus. For the baseline evaluation, a no-POS (no part-of-speech) version of each lexicon was constructed and used. The Hu & Liu and NRC lexicons already do not have part-of-speech tagging. For the other lexicons, when there are conflicting sentiment values assigned to the same word for different parts-of-speech, we used the following order of preference: adjective, adverb, verb and noun. Thus, for a particular word, we prefer the sentiment value assigned to the adjective use of the word, over verb and noun. This is based on the assumption that adjectives are the main bearers of sentiment. Over half of the adjectives and adverbs in the lexicon have positive or negative sentiment, compared to 27% of verbs and 24% of nouns (based on statistics from the WKWSCI lexicon).

Table 3. Evaluation of Baseline SVM and Naïve Bayes models, with negation handling.

	SVM (tf*idf weighting)		Naïve Bayes (tf*idf weighting)	
	Positive	Negative	Positive	Negative
Precision	0.837	0.842	0.829	0.840
Recall	0.843	0.835	0.843	0.827
F1 Score	0.840	0.838	0.836	0.833
Accuracy	0.839 (without negation handling: 0.811)		0.834 (without negation handling: 0.822)	

4.3.1. Experiment 1a: Baseline machine learning method using bag-of-words.

For this experiment, a stoplist of 87 words (with document frequency below 5) was used. Negated terms were converted to features. For example, "not bad" and "not so bad" were represented as *not_bad*. Negated term features with document frequency less than 5 were dropped.

Support Vector Machine (SVM) and Naïve Bayes categorizers were built using the training dataset. Two weighting schemes were used: tf (term frequency) and tf*idf (tf/log document frequency). The results were slightly better for the tf*idf weighting, shown in Table 3. It can be seen that the results were about the same for the SVM and Naïve Bayes models, with accuracy rates of 83% to 84%. Negation handling increased the accuracy from 81% to 84% for the SVM model, and from 82% to 83% for the Naïve Bayes method.

4.3.2. Experiment 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 score is then normalized by the number of words in the review (after removing stopwords). We found that the normalized score consistently gave better results than the unnormalized score. Negated sentiment words were handled as described earlier, i.e. their polarity was reversed. The reviews were then ranked in decreasing score, and the top half of the reviews were categorized as positive, and the bottom half negative. As other authors [e.g., 6] have noted, lexicon-based sentiment scoring is slightly biased towards positive valence. If top half of the scores are categorized as positive polarity, the threshold value separating positive and negative predictions is generally between 0.03 and 0.05 normalized sentiment score.

Table 4. Accuracy of document-level sentiment categorization using normalized baseline scoring method.

LEXICON	Accuracy Without negation handling		Accuracy With negation handling	
	Training set	Test set	Training set	Test set
WKWSCI	0.717	0.709	0.742	0.734
Hu & Liu	0.721	0.719	0.757	0.757
MPQA	0.690	0.699	0.732	0.735
SO-CAL	0.700	0.701	0.741	0.736
General Inquirer	0.642	0.653	0.676	0.688
NRC	0.645	0.646	0.682	0.670
SentiWordNet (weighted average version)	0.605	0.615	0.649	0.645

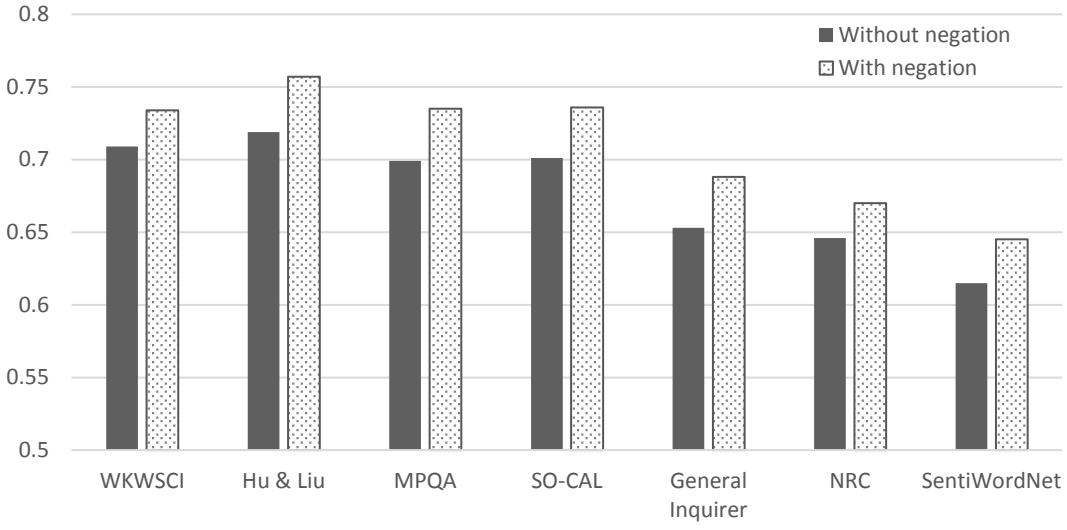

Figure 1. Accuracy of document-level sentiment categorization using normalized baseline scoring method, with and without negation handling

The results are given in Table 4, and presented in a bar chart in Figure 1. As no training is involved in this method, it is not necessary to divide the corpus into training and test set. However, to facilitate comparison with subsequent experiments, we shall focus on the results obtained with the test set.

Hu & Liu lexicon obtained the best accuracy rate of 0.757. WKWSCI, MPQA and SO-CAL had similar results of about 0.735, which was significantly worse than the result of Hu & Liu (α=0.05). General Inquirer, NRC and SentiWordNet obtained weak results of below 70%. SentiWordNet obtained the worst accuracy of 65%, and were dropped from subsequent experiments.

4.3.3. Experiment 2: Lexicon-based method but using logistic regression to determine the weights for different categories of words.

Instead of just counting the number of positive and negative words, different weights can be assigned to different categories of words. However, this requires machine learning on a training set. We used stepwise logistic regression analysis (in the SPSS statistical package) to determine the weights. Each word category is a combination of part-of-speech and sentiment strength (i.e. very positive, positive, slightly positive, slightly negative, negative and very negative). In the logistic regression analysis, each word category is treated as a variable whose value is the number of words of that category found in the document, normalized by dividing by the length of the review. A logistic regression model outputs an estimate of the probability that the document is positive in sentiment. In the experiment, a threshold of 50% probability is used, that is, a document is categorized as *positive* if the estimated probability is 50% or above, and *negative* if the probability is below 50%.

Selected output from the stepwise logistic regression analysis for WKWSCI are given in Tables 5 and 6. Table 5 indicates how significant individual variables in WKWSCI are in predicting the document polarity. It is clear that the normalized baseline score is the best predictor of document polarity. Other strong predictors are: the unnormalized baseline score, number of positive and very positive adjectives/adverbs, and number of negative and very negative adjectives/adverbs. Number of negated positive words is significant, but not negated negative words. Number of positive and number of negative verbs are significant. Number of positive nouns is significant, but not negative nouns. The final regression model shown in Table 6 shows that a combination of normalized and unnormalized baseline score is needed. It also indicates that negated positive words deserve a stronger weight: they count for more than the negative words in the baseline formula. However, negative nouns count for less than other negative words.

Table 5. Stepwise logistic regression output at Step 0 for WKWSCI lexicon: Significance of individual variables.

Variable	Wald Score	P-value
norm_verb_positive1	23.6	0.000
norm_negated_positive23	110.2	0.000
norm_positive1	174.2	0.000
norm_noun_negative1	11.5	0.001
norm_adj&adv_positive1	58.1	0.000
norm_adj&adv_negative1	102.5	0.000
norm_noun_positive23	93.4	0.000
norm_negated_negative23	0.0	0.863
norm_negated_positive1	98.2	0.000
norm_negated_negative1	3.0	0.084
norm_adj&adv_negative23	521.8	0.000
norm_noun_positive1	10.3	0.001
norm_verb_positive23	213.1	0.000
norm_noun_negative23	1.0	0.325
norm_positive23	1123.3	0.000
norm_negative23	441.0	0.000
norm_negative1	279.8	0.000
norm_verb_negative23	123.0	0.000
norm_verb_negative1	165.9	0.000
norm_adj/adv_positive23	1036.8	0.000
unnormalized_baseline	1303.8	0.000
normalized_baseline	1822.3	0.000

Similar insights are obtained from the regression output for the SO-CAL lexicon (Table 7), except that nouns with negative polarity were found to be significant predictors here. In addition, the number of very positive and number of very negative verbs are not significant.

The evaluation results for the logistic regression models are given in Table 8, and presented in a bar chart in Figure 2. Hu & Liu, WKWSCI, MPQA and SO-CAL obtained accuracy rates of 75% to 77% (with no significant differences among them). Using logistic regression to determine the weights to use for different categories of words produced only a small improvement in accuracy of 0.01 to 0.02. SO-CAL obtained a more substantial improvement (from 0.736 to 0.770) possibly because of the more refined 11-point scale sentiment scoring.

These accuracy rates are still worse than those obtained by bag-of-words machine learning models that 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 necessary as the baseline scoring method still gives reasonable results of around 75%.

Table 6. Final logistic regression model for WKWSCI lexicon.

Variable	B	S.E.	Wald	df	P-value
unnormalized_baseline	.121	.012	106.7	1	0.000
normalized_baseline	6.902	.793	75.8	1	0.000
norm_negated_positive23	-17.294	3.179	29.6	1	0.000
norm_negated_positive1	-14.355	3.067	21.9	1	0.000
norm_noun_negative23	36.366	3.083	139.1	1	0.000
norm_positive23	21.010	2.160	94.6	1	0.000
norm_negative23	-20.777	1.705	148.5	1	0.000
norm_verb_negative1	-10.272	2.255	20.7	1	0.000
norm_adj&adv_positive23	-10.068	2.163	21.7	1	0.000
Constant	-.910	.058	244.9	1	0.000

Table 7. Final logistic regression model for SO-CAL lexicon.

Variable	B	S.E.	Wald	df	P-value
unnormalized_baseline	.091	.009	100.7	1	0.000
normalized_baseline	10.159	.659	237.5	1	0.000
norm_adj&adv_negative2	-12.072	2.190	30.4	1	0.000
norm_negated_positive3	-14.420	4.471	10.4	1	0.001
norm_negated_negative1	-18.981	4.260	19.9	1	0.000
norm_positive45	22.567	1.511	223.0	1	0.000
norm_verb_negative3	18.994	2.761	47.3	1	0.000
norm_verb_negative1	-10.211	3.261	9.8	1	0.002
norm_verb_positive1	11.426	2.396	22.7	1	0.000
norm_noun_negative3	19.254	4.181	21.2	1	0.000
norm_positive1	-8.252	1.197	47.5	1	0.000
norm_negative45	-27.036	3.666	54.4	1	0.000
norm_positive2	-3.711	1.382	7.2	1	0.007
norm_adj&adv_positive1	5.393	1.413	14.6	1	0.000
norm_adj&adv_positive2	8.569	1.791	22.9	1	0.000
norm_negative1	12.503	1.180	112.2	1	0.000
Constant	-1.188	.072	270.3	1	0.000

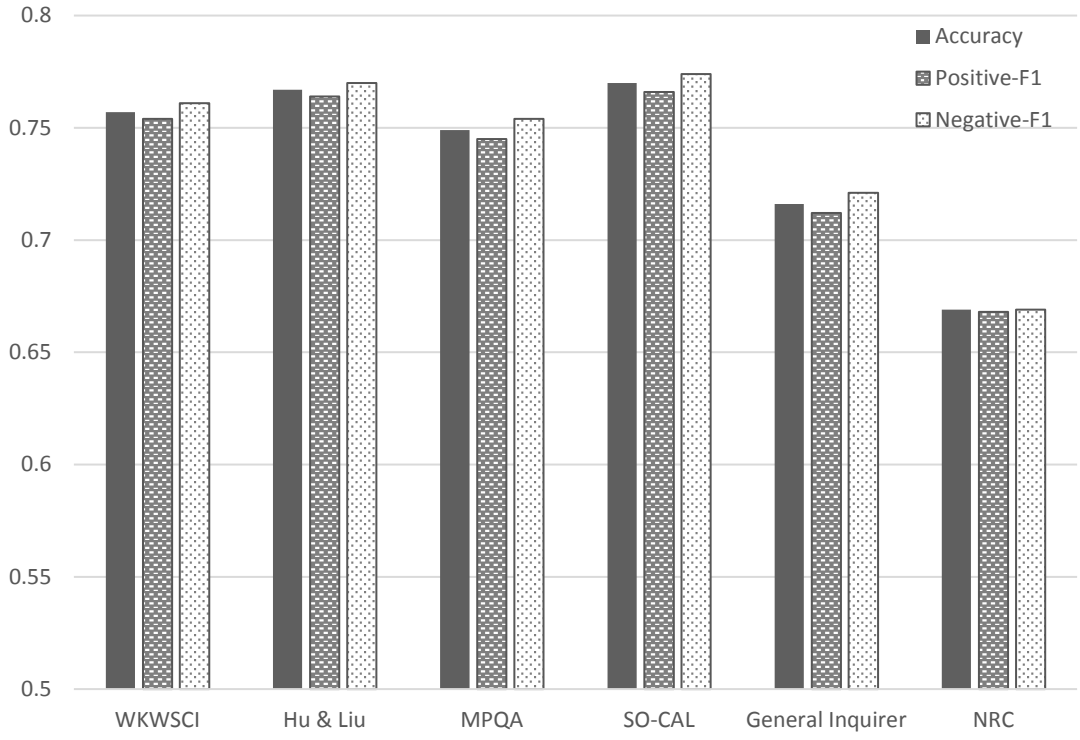

Figure 2. Accuracy and F1 scores for positive and negative reviews, for document-level sentiment categorization, using the logistic regression models.

Table 8. Evaluation results of document-level sentiment categorization on the test set, using the logistic regression models.

LEXICON	Accuracy	Positive reviews			Negative reviews		
		Precision	Recall	F1-score	Precision	Recall	F1-score
WKWSCI	0.757	0.765	0.743	0.754	0.750	0.772	0.761
Hu & Liu	0.767	0.774	0.754	0.764	0.760	0.780	0.770
MPQA	0.749	0.759	0.732	0.745	0.741	0.767	0.754
SO-CAL	0.770	0.779	0.753	0.766	0.761	0.787	0.774
General Inquirer	0.716	0.724	0.700	0.712	0.709	0.733	0.721
NRC	0.669	0.669	0.667	0.668	0.668	0.670	0.669

4.4. Sentence-level sentiment categorization of product reviews

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. The dataset will be made available for public access on the university's institutional repository.[9]

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.

The following experiments were carried out:

- Experiment 3: Baseline lexicon-based method for sentence categorization
- Experiment 4: Lexicon-based method but using logistic regression to determine the weights for different categories of words.

4.4.1. Experiment 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 no part-of-speech versions of the lexicons were used. As many sentences obtained a sentiment score of 0, mainly because no word in the sentence found a match in the lexicon, a sentiment score of 0 was considered to be an incorrect prediction. In other words, the lexicons were penalized for not matching any word in the sentence. Sentences with sentiment score greater than 0 were categorized as *positive*, and sentiment scores below 0 categorized as *negative*.

The accuracy of the sentence-level sentiment categorization is summarized in Table 9. Hu & Liu, WKWSCI, MPQA and SO-CAL obtained similar accuracy rates of about 60%. 7% to 9% of the sentences obtained a sentiment score of 0, for these four lexicons.

4.4.2. Experiment 4: Lexicon-based method but using logistic regression to determine the weights for different categories of words.

Logistic regression was used to determine the weights to use to combine the normalized baseline score with the un-normalized baseline score. Preliminary experiments had indicated that determining the weights for different categories of words (as was done for the document-level sentiment categorization) did not improve accuracy. This was because each sentence had only a small number of sentiment lexicon matches and so the dataset was very sparse, with most word categories having 0 values.

The evaluation results are given in Table 9, and presented in a bar chart in Figure 3. Hu & Liu obtained the highest accuracy of 78%. WKWSCI, MPQA and SO-CAL obtained similar accuracy of about 75% (significantly worse than Hu & Liu at the 0.05 level). Instead of just counting positive and negative words, we also tried using the sentiment strength scores of WKWSCI and SO-CAL (summing up the positive and negative sentiment values of sentiment-bearing words in the sentence). This improved the accuracy to the same level as that obtained by the Hu & Liu lexicon without sentiment strength values.

For comparison, a machine-learning model using the Naïve Bayes method obtained an accuracy rate of 75%. It is surprising that the machine-learning model did as well as the lexicon-based method, given the small training set. The Naïve Bayes method did better than SVM, confirming the findings of Wang and Manning [4] that the Naïve Bayes method is more effective for short reviews, whereas SVM is more effective for full-length reviews.

Table 9. Evaluation results of sentence-level sentiment categorization on the test set, using normalized baseline and logistic regression models.

LEXICON	Using normalized baseline score	Using score from logistic regression model						
			Positive reviews			Negative reviews		
	Accuracy	Accuracy	Precision	Recall	F1-score	Precision	Recall	F1-score
WKWSCI	0.595 (135 sentences with score of 0)	0.752	0.749	0.718	0.733	0.756	0.784	0.770
Hu & Liu	0.594 (160)	0.777	0.786	0.729	0.756	0.772	0.822	0.796
MPQA	0.602 (127)	0.745	0.723	0.748	0.735	0.767	0.743	0.755
SO-CAL	0.592 (135)	0.747	0.735	0.729	0.732	0.759	0.764	0.761
General Inquirer	0.537 (156)	0.676	0.693	0.569	0.625	0.667	0.774	0.716
NRC	0.451 (204)	0.651	0.658	0.550	0.599	0.648	0.743	0.692
SVM model (using tf weighting)	-	0.727	0.713	0.710	0.711	0.741	0.743	0.742
Naïve Bayes model (using tf weighting)	-	0.754	0.744	0.733	0.738	0.764	0.774	0.769

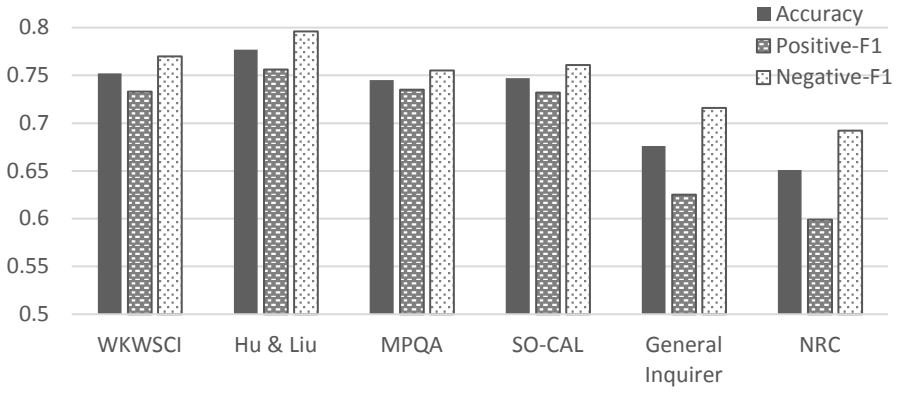

Figure 3. Accuracy and F1 scores for positive and negative reviews, for sentence-level sentiment categorization, using the logistic regression models.

4.5. Sentiment analysis of news headlines

The following experiments were carried out on the news headlines corpus using the sentiment lexicons:

- Experiment 5: Sentiment categorization of news headlines into positive or negative polarity
- Experiment 6: Prediction of the sentiment strength of the news headlines.

4.5.1. Experiment 5: Sentiment categorization of news headlines using a lexicon-based method

For this experiment, a subset of the news headlines corpus containing only clearly positive headlines (with scores of 50 to 100) and clearly negative headlines (with scores -50 to -100) were used. There are 410 such headlines, comprising 155 positive and 255 negative headlines.

The baseline scoring method (*number of positive words - number of negative words*) was used to calculate the sentiment score for each headline. Headlines with sentiment score greater than 0 were categorized as positive, and sentiment scores below 0 were categorized as negative. A score of 0 (usually because no word in the headline is found in the lexicon) was evaluated as incorrect prediction. So, a lexicon is penalized for not having any matches in the headline.

The evaluation results are given in Table 10, and presented visually in Figure 4. The WKWSCI lexicon obtained the best accuracy of 69%, which is substantially better than the other lexicons. General Inquirer and NRC, which had not fared well with the product review corpus, did well with the news headlines corpus, obtaining accuracy rates of 64% and 65% respectively. The Hu & Liu lexicon did creditably (62% accuracy), even though it was constructed for the product reviews domain.

Taboada et al. [6] used the same corpus in their evaluation of the SO-CAL program, but categorized the headlines that obtained a 0 sentiment score as *negative* (on the basis that negative headlines are in the majority). If we categorize the headlines with 0 score as *negative*, then WKWSCI obtains a higher accuracy of 75.6%, and SO-CAL 67.8%.

Table 10. Evaluation results of news headlines sentiment categorization using the normalized baseline method.

LEXICON	Accuracy	No. (%) of headlines with no word match in the lexicon
WKWSCI	0.692	85 (21%)
Hu & Liu	0.624	130 (32%)
MPQA	0.582	120 (29%)
SO-CAL	0.570	124 (30%)
General Inquirer	0.641	83 (20%)
NRC	0.649	87 (21%)

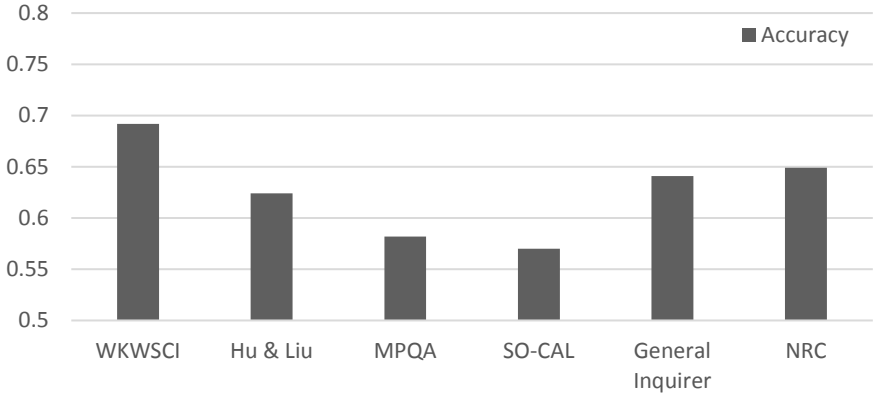

Figure 4. Accuracy of news headlines sentiment categorization using the normalized baseline method.

4.5.2. Experiment 6: Sentiment strength prediction using WKWSCI and SO-CAL

WKWSCI and SO-CAL were the only lexicons in this study that assign a sentiment strength value to each word. We wanted to find out how accurately the lexicons can predict the actual sentiment scores assigned to the news headlines by human coders. We applied both lexicons to the full news headlines dataset, to calculate the sentiment scores using the normalized baseline scoring method.

As evaluation measure, we calculated the Pearson r correlation between the sentiment scores assigned by human coders and the sentiment scores assigned using each lexicon. The results are given in Table 11. The correlation between human scoring and WKWSCI was a moderate 0.57, which is significantly better (α=0.01) than the correlation with SO-CAL lexicon.

For a head to head comparison between WKWSCI and SO-CAL, we dropped the news headlines that had no matching word with either WKWSCI or SO-CAL (or both). WKWSCI obtained a significantly better correlation of 0.66, compared to 0.60 for SO-CAL. The correlation between WKWSCI and SO-CAL scores were 0.80. We carried out a linear regression analysis that found that combining WKWSCI and SO-CAL sentiment scores can significantly improve the sentiment score prediction.

In the SemEval-2007 Workshop where the corpus was first used, the best performance obtained by the five participating systems was a correlation of 0.48 [15]. In developing the gold standard, Strapparava and Mihalcea found that the six coders had an average correlation of 0.78 (taking the correlation between each coder and the average of the other five coders).

Table 11. Correlation between predicted sentiment strength with target values.

LEXICON	Correlation with target values	
	Full dataset	Subset with non-zero scores for both WKWSCI & SO-CAL
WKWSCI	0.568	0.660
SO-CAL	0.479	0.601

5. Error analysis

We examined the false positive and false negative errors for the WKWSCI lexicon using the baseline scoring method. For the product review sentences, there were 22 false negatives and 62 false positives in the test set (excluding sentences with a sentiment score of 0). The two main sources of error are:

- Product-specific attributes that are positive or negative in the context
- Indirect or roundabout expressions of satisfaction or dissatisfaction.

Here are some examples of product-specific attributes that imply a positive/negative sentiment:

Positive sentiment
The nonstick surface seems to be less prone to scratch.
We use these sock snowmobile in -20F temperature and they keep your toes from getting numb when you are outdoors for a long stretch .
The narrow neck means that the fruit will not fall into your glass while you are pouring and the fruit stays put also means it can steep in the water longer.
Negative sentiment
The rubber section on these are so thin that they do not even seal the wine bottle.
It prints nicely but it does not help if it ruins your fax.
The noise level is comparable to an industrial vacuum cleaner.

Domain-specific product attributes are too many to enumerate in a dictionary, and most do not occur often in the corpus. Common sense knowledge and reasoning are needed to determine the sentiment.

Indirect expressions of sentiment are expressed in many ways. We identified the following types of positive comments:

- Domain-specific positive experience: for example, *I cry every time I watch it [a movie]*.
- Evokes envy: *The setup is quick and we are the envy of all beachgoers.*
- A favourite tool that is often used: *This is one of my most used knife in the kitchen.*
- Double negation: *I cannot imagine how anyone would dislike this movie.*
- Product durability: *I purchased this product 2 years back and it is still working without any problem.*
- Good service: *Arrived shortly after placing order.*
- Star rating: *I would have given it five stars but heck nobody's perfect.*

We identified the following types of negative comments:

- Looks good, but not functional: *It is pretty but that is about it.*
- Looks cheap: *It looks and feels like a cheap vase one would buy at Target.*
- Decision to return product: *Right then and there I was ready to send this thing back.*
- Decision not to return product: *It is inexpensive enough that I will not bother to return it*.
- Product not used: *I never wear this top after having purchased it.*
- Comparison with other products: *For the money there are much better systems out there.*
- Exception to good experience: *I have loved every le creuset product I have purchased with the big exception of the whistle teakettle!*
- Decision to give away: *I think I have worn it once and will probably donate it soon.*
- Decision to buy a different product: *We have to buy a different pair of shoes that form to his foot better.*
- Poor service: *The customer service rep told me that it would be taken care of 2 weeks later I got another bill.*
- Star rating: *Do not deserve the one star.*

Other indirect expressions of sentiment are more generic and more common, and can probably be enumerated in a dictionary. Here are some examples that reflect positive sentiment:

… it would be hard to go wrong with this one.
… is expensive but you'll definitely get what you pay for. (This can be positive or negative depending on the context.)
… this is a must see.
… this will spoil you.
… you cannot find a better one than this model ...
… once you try them you never have to worry about … again.
This is the only type of shoe I will ever buy …
It does everything I expect it to do.
Does exactly what it is supposed to do.
Could not have been much easier .
… I intend to buy again ... and again ... and again
It is exactly what I have been looking for.
As you can see the pro outweighs the con by quite a bit.

Video/movie reviews are particularly problematic. They tend to have long complex sentences that often include a summary of the plot, which needs to be excluded from the sentiment analysis. Here are some examples that reflect positive sentiment:

"Badlands" Terrence Malick's 1973 directorial debut an overlooked jewel of a film forcibly moves you and takes an unsavory subject the Charles Starkweather serial killing crime spree up in the vacated badlands of South Dakota and refine it til you posit "how do we get here as a country"?
Leslie Nielsen gave the performance of his life as a psychotically jealous husband who take a notion to bury his wife and her lover neck-deep in sand before the tide rolls in.
Well anyone who thinks that the salsa dancing in this movie is amazing, awesome, good or ok have never actually seen salsa dancin!!

Users sometimes use sarcasm and hyperbole to express negative sentiments:

Thanks for deleting part of my childhood.

Am I supposed to care what happens to him?
It may be number 1 in Germany but it gets only 1 star here.
I recommend this movie to 6 year olds who don't care about the quality of movie.
To get the stench of this swill out of my consciousness I am going to have to watch an actual good movie like "showgirls".
Hopefully they have fixed the problem, but I would not bet my money on it.

6. Conclusion

We have described the characteristics of five major sentiment lexicons in comparison with a new general-purpose sentiment lexicon (WKWSCI Sentiment Lexicon) that we developed using manual coding. WKWSCI, General Inquirer, MPQA and SO-CAL lexicons distinguish between the different parts-of-speech, whereas NRC and Hu & Liu lexicons do not. WKWSCI and SO-CAL also provide sentiment strength values (a 7-point sentiment scale in the case of WKWSCI, and an 11-point scale for SO-CAL). A majority of the coding conflicts between WKWSCI lexicon and the other lexicons involve neutral words in the WKWSCI lexicon being coded positive or negative in the other lexicon. The conflicting polarities are due to words having different sentiments for different word senses and contexts.

We initially included SentiWordNet in the study, but decided to leave it out as it performed poorly in the baseline experiment and its sentiment coding was found to differ substantially from the other lexicons. This is possibly because SentiWordNet assigns different sentiment scores to different word senses, and sense disambiguation is probably needed to get good results from its use.

We evaluated the effectiveness of the six lexicons in a sentiment categorization task on an Amazon product review corpus and a news headlines corpus. A simple method of handling negation was used in this study: if a negation word occurs to the left of a sentiment-bearing word, with up to one word in between, the sentiment polarity is reversed and a weak strength is assigned.

WKWSCI, MPQA, Hu & Liu and SO-CAL lexicons are equally good for product review sentiment categorization, obtaining accuracy rates of 75% to 77% when the appropriate weights for the different categories of sentiment words are determined using logistic regression analysis on a training set. If a training set is not available, the simple-minded formula—number of positive words minus number negative words—gives results that are nearly as good. Better results are obtained by normalizing the score by the length of the review. In this case, the Hu & Liu lexicon gave the best accuracy of 76%. This is not surprising as the lexicon was derived from product review texts and is thus customized to the domain. The SO-CAL lexicon benefitted the most from differentially weighting different categories of words, as it distinguishes between parts-of-speech and uses a refined 11-point scale for scoring the sentiment strength of words. Its accuracy rate improved from 73.6% to 77% with the weighting.

The study has also sought to determine the relative importance of different categories of words in predicting sentiment polarity at the document level. The results of logistic regression analyses indicate that, generally, a combination of normalized and unnormalized baseline score (using the simple-minded formula) is needed. It also indicates that negated positive words deserve a stronger weight: they count for more than negative words and negated negative words. It is well-known that adjectives are important in indicating sentiment. However, verbs and nouns are found to be important too, but positive nouns carry more weight than negative nouns.

A bag-of-words machine-learning model using Support Vector Machine obtained an accuracy of about 84% for document-level sentiment categorization. This probably represents the upper bound of what can be achieved using a lexicon-based method. Our method of handling negation improved the accuracy from 81% to 84%. 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 sentiment categorization of sentences in the product reviews, we combined two versions of the simple-minded formula—with and without normalization—using logistic regression to determine their relative weights. Hu & Liu lexicon obtained the best accuracy of 78%. We were not able to determine the weights to use for different parts-of-speech using logistic regression, as each sentence had only a small number of sentiment lexicon matches.

For sentiment categorization of a news headlines dataset, the WKWSCI lexicon obtained the best accuracy of 69%. For predicting the sentiment strength value (manually assigned by human coders), the values assigned using the WKWSCI lexicon obtained a moderate Pearson r correlation of 0.57 with human assigned sentiment values. If the headlines with no matching word are excluded, the correlation increased to 0.66.

We tentatively recommend that the Hu & Liu lexicon be used for product review texts, and the WKWSCI lexicon be used for non-review texts. Of course, experiments on a wider range of non-review texts are needed to obtain more definite conclusions, and to provide a basis for improving the lexicons for non-review texts.

A limitation of this study is that it made use only of single-word terms. Use of multiword expressions (e.g., "do not buy") is very likely to improve the accuracy of sentiment analysis. Use of intensifiers (e.g., "strongly"), maximisers (e.g., "absolutely"), mitigators (e.g., "slightly") and minimizers (e.g. "scarcely") are also likely to help, especially at the sentence level. If a sentence contains both a positive and a negative word, an intensifier can indicate which sentiment is stronger or more important. It should be pointed out that the evaluation for the SO-CAL lexicon is not entirely fair as it includes a relatively large number of multiword expressions that were not used in this study. The SO-CAL application program also makes use of intensifiers and mitigators to calculate sentiment values. These issues will be examined in the next phase of the project.

Notes

1. Downloaded from: http://www.wjh.harvard.edu/~inquirer/spreadsheet_guide.htm
2. Downloaded from: http://mpqa.cs.pitt.edu/lexicons/subj_lexicon/
3. Downloaded from: http://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html#lexicon
4. Downloaded from: http://saifmohammad.com/WebPages/lexicons.html
5. Obtained from the authors.
6. Downloaded from: http://sentiwordnet.isti.cnr.it/
7. https://www.mturk.com/
8. Downloaded from: http://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html
9. https://repository.ntu.edu.sg/

Funding

This work is funded by Academic Research Fund Tier 1 RGT38/13 from Nanyang Technological University, Singapore.

References

[1] Cortes C and Vapnik V. Support vector networks. *Machine Learning* 1995; 20(3): 273-297.
[2] Vapnik VN and Vapnik V. Statistical Learning Theory. New York: John Wiley and Sons, 1998.
[3] Zhang H. The optimality of naive Bayes. In: *Proceedings of the Seventeenth Florida Artificial Intelligence Research Society Conference*, 2004, pp.562–567. American Association for Artificial Intelligence Press.
[4] Wang S and Manning CD. Baselines and bigrams : simple, good sentiment and topic classification. In: *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, Jeju, Republic of Korea, 2012, pp.90–94. Association for Computational Linguistics.
[5] Stone PJ, Dunphy DC, Smith MS, et al. The General Inquirer: A Computer Approach to Content Analysis. Cambridge, MA: MIT Press, 1966.
[6] Taboada M, Brooke J, Tofiloski M, et al. Lexicon based methods for sentiment analysis. *Computational Linguistics* 2011; 37: 267–307.
[7] Hatzivassiloglou V, Mckeown KR. Predicting the semantic orientation of adjectives. In: *Proceedings of 35th Meeting of the Association for Computational Linguistics*, 1997, pp. 174–181. Association for Computational Linguistics.
[8] Turney P and Littman M. Measuring praise and criticism: inference of semantic orientation from association. *ACM Transactions on Information Systems* 2003, 21(4): 315–346.
[9] Esuli A and Sebastiani F. SentiWordNet: a publicly available lexical resource for opinion mining. In: *Proceedings of 5th International Conference on Language Resources and Evaluation*, 2006, pp.417–422.
[10] Das A and Gambäck B. Sentimantics: conceptual spaces for lexical sentiment polarity representation with contextuality. In: *Proceedings of the 3rd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis*, Jeju, Republic of Korea, 2012, pp.38–46. Association for Computational Linguistics
[11] Thelwall M and Buckley K. Topic-based sentiment analysis for the social web: The role of mood and issue-related words. *Journal of the American Society for Information Science and Technology* 2013; 64: 1608–1617.
[12] Thet TT, Na JC and Khoo CSG. Aspect based sentiment analysis of movie reviews on discussion boards. *Journal of Information Science*, 2010; 36(6): 823-848.
[13] Wiebe J, Wilson T and Cardie C. Annotating expressions of opinions and emotions in language. *Language Resource Evaluation* 2005; 39: 165–210.
[14] Khoo CSG, Nourbakhsh A and Na JC. Sentiment analysis of news text: a case study of appraisal theory. *Online Information Review* 2012, 36(6): 858-878.
[15] Strapparava C and Mihalcea R. SemEval-2007 Task 14: Affective Text. In: *Proceedings of the 4th International Workshop on Semantic Evaluations*, 2007, pp. 70–74. Prague: Association for Computational Linguistics,
[16] *12 Dicts Introduction*. Retrieved from http://wordlist.aspell.net/12dicts-readme/

[17] Khoo CSG, Johnkhan SB & Na JC. Evaluation of a general-purpose sentiment lexicon on a product review corpus. In: *Proceedings of 17th International Conference on Asia-Pacific Digital Libraries*. Seoul, Republic of Korea, 2015, pp. 82–93. Berlin: Springer.

[18] Mohammed SM and Turney DP. Crowdsourcing a word-emotion association lexicon. *Computational Intelligence* 2013, 29(3): 436-465

[19] Hong Y, Kwak H, Baek Y, et al. Tower of babel: a crowdsourcing game building sentiment lexicons for resource-scarce languages. In: *WWW 2013 Companion*, Rio de Janeiro, Brazil, 2013. New York: ACM.

[20] Thisone CC, Ghasemi A and Faltings B. Sentiment analysis using a novel human computation game. In: *Proceedings of the 3rd Workshop on the People's Web Meets NLP*, Jeju, Republic of Korea, 8-14 July 2012, 2012, pp. 1–9. Association for Computational Linguistics

[21] Kamps J, Marx M, Mokken RJ, et al. Using wordnet to measure semantic orientation of adjectives. In: *Proceedings of the Language Resources and Evaluation*. 2004, pp. 1115-1118. Paris: European Language Resources Association.

[22] Church KW and Hanks P. Word association norms, mutual information, and lexicography. *Computational Linguistics* 1990; 16(1): 22-29.

[23] Glavaš G, Šnajder J and Bašić BD. Experiments on hybrid corpus based sentiment lexicon acquisition. In: *Proceedings of the Workshop on Innovative Hybrid Approaches to the Processing of Textual Data*, Avignon, France, 2012, pp. 1–9. Association for Computational Linguistics.

[24] Li F, Pan SJ, Jin O, et al. Cross-domain co-extraction of sentiment and topic lexicons. In: *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, Jeju, Republic of Korea, 8-14 July 2012, pp. 410–419. Association for Computational Linguistics.

[25] Qiu G, Liu B, Bu J, et al. Expanding domain sentiment lexicon through double propagation. In *Proceedings of the 21st International Joint Conference on Artificial Intelligence*, 2009, vol. 9, pp. 1199–1204. Morgan Kaufmann Publishers Inc.

[26] Bahrainian SA, Liwicki M and Dengel A. Fuzzy Subjective sentiment phrases: a context sensitive and self-maintaining sentiment lexicon. In: *Proceedings of the 2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)*, 2014, Vol. 1, pp.361-368. IEEE Computer Society.

[27] Yates A, Goharian N and Yee WG. Semi-supervised probabilistic sentiment analysis: merging labeled sentences with unlabeled reviews to identify sentiment. In: *Proceedings of the American Society for Information Science and Technology*, Montreal, Quebec, Canada, 2013, November 1-6, 50(1), pp.1-10.

[28] Xu X, Tan S, Liu Y, et al. Towards jointly extracting aspects and aspect-specific sentiment knowledge. In: *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, Maui, HI, USA, 2012, Oct 29, pp. 1895-1899. New York: ACM.

[29] Blei DM, Ng AY and Jordan MI. Latent dirichlet allocation. *Journal of Machine Learning Research* 2003; 3: 993-1022.

[30] Thelwall M, Buckley K and Paltoglou G. Sentiment strength detection for the social web. *Journal of the American Society for Information Science and Technology* 2012; 63(1): 163-73.

[31] Devitt A and Ahmad K. Is there a language of sentiment? An analysis of lexical resources for sentiment analysis. *Language Resources and Evaluation* 2013 Jun 1; 47(2): 475-511.

[32] Thorndike EL and Lorge I. *The Teacher's Word Book of 30,000 Words*. 4th ed. New York: Columbia University Press, 1963.

[33] Lasswell HD and Namenwirth JZ. *The Lasswell Value Dictionary*. Vols. 1–3. New Haven: Yale University, 1968.

[34] *Origins of the General Inquirer Marker Categories*. Retrieved from *http://www.wjh.harvard.edu/~inquirer/kellystone2.htm*

[35] Riloff E and Wiebe J. Learning extraction patterns for subjective expressions. In: *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, 2003, pp. 105–112. Association for Computational Linguistics.

[36] Hu M and Liu B. Mining and summarizing customer reviews. In: *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2004, pp.168-177. New York: ACM.

[37] Bernard J (ed.). *The Macquarie Thesaurus*. Sydney: Macquarie Library, 1986.

[38] Brants T and Franz A. *Web 1T 5-Gram Version 1*. Linguistic Data Consortium, 2006.

[39] Strapparava C and Valitutti A. Wordnet-Affect: An affective extension of WordNet. In: *Proceedings of the 4th International Conference on Language Resources and Evaluation LREC-2004*, Lisbon, Portugal, 2004, pp. 1083–1086.

[40] Pang B, Lee L and Vaithyanathan, S. 2002. Thumbs up? Sentiment classification using machine learning techniques. In: *Proceedings of the Conference on Empirical Methods in NLP*, Philadelphia, PA, pp. 79–86.

[41] Jindal N and Liu B. Opinion spam and analysis. In: *Proceedings of the 2008 International Conference on Web Search and Data Mining*, 2008, pp. 219-230. New York: ACM.

[42] Manning CD, Surdeanu M, Bauer J, et al. The Stanford CoreNLP Natural Language Processing Toolkit. In: *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 2014, Jun 23, pp. 55-60.

[43] Bird S, Loper E and Klein E. *Natural Language Processing with Python.* O'Reilly Media Inc., 2009.

[44] Na JC, Khoo, CSG and Wu PHJ. Use of negation phrases in automatic sentiment classification of product reviews. *Library Collections, Acquisitions & Technical Services*, 2005; 29(2): 180-191.