Proceedings of the Twenty-Ninth International
Florida Artificial Intelligence Research Society Conference

Sentiment Classification Using Negation as a Proxy for Negative Sentiment

Bruno Ohana and **Brendan Tierney** and **Sarah Jane Delany**
Dublin Institute of Technology, Kevin Street
Dublin 8, Ireland

Abstract

We explore the relationship between negated text and negative sentiment in the task of sentiment classification. We propose a novel adjustment factor based on negation occurrences as a proxy for negative sentiment that can be applied to lexicon-based classifiers equipped with a negation detection pre-processing step. We performed an experiment on a multi-domain customer reviews dataset obtaining accuracy improvements over a baseline, and we further improved our results using out-of-domain data to calibrate the adjustment factor. We see future work possibilities in exploring negation detection refinements, and expanding the experiment to a broader spectrum of opinionated discourse, beyond that of customer reviews.

1 Introduction

Sentiment Classification is the task of predicting the sentiment orientation of subjective text as positive or negative. With the ever increasing volume of user generated content available for mining in blogs, online product reviews and forums, this task has received considerable research attention in the past decade. In broad terms, sentiment classification approaches can be grouped into: (i) using a supervised learning technique where a training set of documents is transformed into a suitable data representation for a machine learning algorithm; (ii) taking advantage of pre-existing language resources that facilitate the extraction of sentiment information via natural language techniques, and (iii) a combination of the above.

One language resource popular in sentiment classification tasks is the *sentiment lexicon*: a database that maps words and expressions to sentiment information, typically encoded as a numerical score. Sentiment lexicons attempt to capture pre-existing knowledge on a word's sentiment (its *prior polarity*) obtained from human annotation or an automated approach that expands a set of seed words using lexical resources or corpora. The information from the lexicon can be used as additional features in a supervised learning classifier, or alternatively a classifier can determine document sentiment by evaluating the aggregate sentiment of words found. The latter technique lends itself well to natural language methods that identify clues relevant to document sentiment, and one important such clue to sentiment analysis is whether a passage has been negated.

Negation detection has two main tasks: determining if negation has occurred in a given passage, and what is the region of text affected by it, or its *scope*. Negation realizations in language can vary considerably as illustrated in the examples below: scope can be explicitly demarcated with well known words (I), or on terms semantically negated to absence, rejection or failure (II). Negating expressions however may not always imply negation and need further disambiguation (III).

(I) "I did *not* like the food *but* the hotel bed was great."
(II) "I *fail to* see how this movie got such good reviews."
(III) "*Not only* was the hotel dirty, it was also noisy."

Negation is a frequent phenomenon in both formal and informal text, reported in various studies as surveyed by Morante and Sporleder (2012), including opinionated text, where an annotation task of customer reviews reported 19% of sentences containing negation (Councill, McDonald, and Velikovich 2010). A negated passage will affect the sentiment of words within its scope, making automated negation detection methods an area of considerable interest to sentiment classification.

Underpinning such methods is the treatment of negation as implying logical inversion of meaning, which is used by a classification algorithm to determine when the polarity of sentiment words should be inverted. However negation occurrence in natural language can be more nuanced: a study from Potts (2011) evaluated the effect of negating terms and their correlation to negative sentiment in opinionated text and revealed explicit negation markers to occur more frequently in negative sentiment text. This apparent preference can be explored by sentiment classifiers that employ negation detection in their evaluation, and is the main focus of this study.

In this paper we investigate the relationship between negative sentiment and negation in sentiment classification. Using a lexicon-based classifier and a rule-based negation detection pre-processing step, we introduce an adjustment to negative sentiment based on negated passages and investigate whether out-of-domain data can be used to dynamically set this adjustment factor. The rest of this paper is organized as follows: Section 2 discusses related work in sentiment classification, negation detection and its role in sentiment analysis; Section 3 investigates the negative content of negating words in commonly used sentiment lexicons in the literature. In Section 4 we present an experiment evaluating the effects of negation on a lexicon-based sentiment classification task using multi-domain customer reviews dataset, and discuss results with respect to previous research in the literature. Section 5 presents final remarks and future work opportunities.

2 Related Work

The goal of sentiment classification is to determine what, if any, is the sentiment orientation of a given input text. In particular we are interested in the overall sentiment conveyed at document level, and apply the assumptions stated in the *document-level sentiment classification* task from (Liu and Zhang 2012) where sentiment within the document comes from a single opinion holder and refers to a single entity, for example a review on a particular product. The most common characterization of a sentiment classification task is that of a binary classification problem with positive and negative classes, but it can also be modelled as regression or multi-class classification problem such as film reviews feedback in a numeric scale (Pang and Lee 2005) or satisfaction scores on travel destinations (Baccianella, Esuli, and Sebastiani 2009).

Approaches to document sentiment classification in the literature use a combination of supervised learning and methods that take advantage of a pre-existing resource to extract sentiment information via natural language techniques. Supervised learning methods have been extensively studied in this task: early work from Pang, Lee, and Vaithyanathan (2002) presents a series of experiments using different classifiers and n-gram word vectors as features on a film review dataset. Later work from Cui, Mittal, and Datar (2006) shows that higher order n-gram vectors can obtain good results when significantly larger datasets are used (over 300k product reviews). More recent studies obtained improvements by experimenting with adapted *tf-idf* feature weight schemes (Paltoglou and Thelwall 2010). Approaches that extend the feature set with document statistics are seen in (Dave, Lawrence, and Pennock 2003) and in (Abbasi, Chen, and Salem 2008) an additional feature selection preprocessing step also yields improved classifier performance. The feature vector is extended in (Whitelaw, Garg, and Argamon 2005) with features relevant to sentiment classification based upon the appraisal language framework from Martin and White (2005). More recent approaches exploring features extracted from multi-modal data sources can be seen for example in (Poria, Cambria, and Gelbukh 2015). We refer the reader to surveys in (Pang and Lee 2008) and (Liu and Zhang 2012) for a deeper discussion on supervised sentiment classification methods.

Sentiment lexicons are databases that store a-priori sentiment information of words and expressions. Lexicons have been created in the literature by combinations of techniques that include manual annotation (Taboada et al. 2011; Wilson, Wiebe, and Hoffmann 2005), crowdsourcing (Mohammad and Turney 2013) and algorithmic expansion from a set of seed words exploring relationships encoded in existing resources such as the SentiWordNet lexicon (Baccianella, Esuli, and Sebastiani 2010), or in linguistic patterns extracted from a corpus (Hatzivassiloglou and McKeown 1997; Goyal and Daumé III 2011).

The information from sentiment lexicons can be used to classify document sentiment with a term counting and aggregation strategy that dispenses with training data, and preserves the original document structure, thus making such methods good candidates for exploring natural language patterns that indicate or modify sentiment. The work of (Kennedy and Inkpen 2006) uses a manually built lexicon and a term-counting method to determine document sentiment, and also employs the detection of intensifier and diminisher terms (*very, little*, etc.). Multi-domain experiments using similar approaches have been performed in (Taboada et al. 2011; Ding, Liu, and Yu 2008). Lexicon-based approaches can be combined with other approaches to form more robust classification frameworks: lexicons can be used as features in supervised learning methods as seen in the use of the SentiWordNet in (Denecke 2009; Gezici et al. 2012; Kiritchenko, Zhu, and Mohammad 2014); the work of (Poria et al. 2014) introduces a framework using a database of affective concepts, linguistic rules and machine learning applied to sentence-level sentiment classification.

Of particular interest to us is the treatment of negated passages when classifying sentiment. The early work of Pang, Lee, and Vaithyanathan (Pang, Lee, and Vaithyanathan 2002) performs a negation detection pre-processing step that searches for explicit negating words and prefixes the words following it with an artificial *NOT* tag. This modified text is used as input to a bag-of-words classifier. The study does not present comparative data but reports that *"removing the negation tag had a negligible, but on average slightly harmful effect on performance"*. As observed by (Wiegand et al. 2010), some negation patterns can be captured by higher order n-gram models, for example in bi-grams such as *not interesting*. This can partially explain the good performance of in-domain supervised learning techniques on experiments that dispense with negation tags.

In lexicon-based classifiers that use term counting strategies, a common approach is to have word polarity inverted when found within a negated passage (Kennedy and Inkpen 2006; Councill, McDonald, and Velikovich 2010; Ding, Liu, and Yu 2008). In (Taboada et al. 2011) word polarity is instead shifted by a fixed value, reflecting the intuition that the polarity of negated words do not necessarily carry the same intensity as the original word. For example *not excellent* would not indicate negative sentiment, but rather an attenuation in the strength of the word *excellent*. In (Kiritchenko, Zhu, and Mohammad 2014) this approach is further refined by constructing separate lexicons for words occurring in an affirmative or a negated context. Polarity shift and attenuation are determined by observed frequencies in a training set of labelled tweet messages. The authors report that 76% of the positive words reverse their polarity when inside a negated context, while 82% of the nega-

Table 1: Negating words in the General Inquirer (GI), MPQA and SentiWordNet 3.0 (SWN3) lexicons.

	GI	MPQA	SWN3	
Word	Negativ	Prior Polarity	Positive Score	Negative Score
not	-	-	0.375	0.625
none	-	-	0.375	0.625
no	-	-	0.375	0.625
never	-	-	0.125	0.625
nobody	-	-	0.0	0.0
nothing	-	-	0.25	0.25
neither	-	-	0.0	0.25
nor	-	-	-	
nowhere	-	-	0.0	0.125
without	-	-	-	
lack	✓	negative	0.125	0.125
hardly	-	negative	0.125	0.25

tive words retain the same polarity but shift their sentiment scores.

3 Negating Words and Sentiment Lexicons

Using the set of commonly used negation words presented in (Councill, McDonald, and Velikovich 2010) we inspected how they are encoded in three popular sentiment lexicons in the literature, with results in Table 1. The General Inquirer (GI) lexicon (Stone et al. 1966) is a manually compiled lexical resource containing linguistic annotations for words in English, including for positive and negative polarity. In this lexicon only one of the negating terms appears as carrying negative sentiment (*Negativ* tag). Interestingly, some negating words (*Negate* tag) are annotated with other negative-biased tags (hostile, weak), and GI's documentation[1] suggests a possible link to negative sentiment:

> Negate - has 217 words that refer to reversal or negation, including about 20 "dis" words, 40 "in" words, and 100 "un" words, as well as several senses of the word "no" itself; *generally signals a downside view*.

The MPQA Subjectivity Lexicon (Wilson, Wiebe, and Hoffmann 2005) is derived from prior polarity annotations from a corpus of subjective text, coupled with sentiment word lists found on other resources, including GI. In this lexicon two negating words are annotated with prior sentiment polarity. SentiWordNet 3.0 (SWN3) (Baccianella, Esuli, and Sebastiani 2010) is an automated lexicon generated from a set of seed words and an expansion algorithm based on word relationships and glosses encoded in the WordNet database. In this lexicon 7 of the 12 words carry negative sentiment, with the remainder being either neutral or not present[2]. Each word sense in SentiWordNet is associated with a tuple of positive and negative polarity values, and we present the highest of each polarity when multiple senses are found.

4 Application to Sentiment Classification

The lexicons surveyed in Table 1 indicate, with some exceptions, a preferred interpretation of sentiment-neutral logical negation in their encoding of negating words. This encouraged us to investigate whether a consistent treatment of negative sentiment in negated text can improve the performance of sentiment classification tasks, given the observations from (Potts 2011).

We start by building a baseline experiment that evaluates the effectiveness of negation detection using a lexicon-based classifier. Following similar methods in the literature (Taboada et al. 2011) our algorithm tokenizes and tags an input document for part-of-speech (using the Stanford Part-of-speech Tagger[3]), and queries sentiment information in each word from an input sentiment lexicon. Scores for document-wide positive and negative polarity are calculated as the sum of individual word scores, and the classification decision is based on the class with the highest score. In SWN3, each word sense is assigned a numerical tuple indicating positive and negative sentiment between 0 and 1. The MPQA and GI lexicons provide only textual annotations, which we convert into numerical values. MPQA also annotates polarity strength as *strong* and *weak*, which we represent respectively with scores of 1.0 and 0.5. We disambiguate word sense by part-of-speech and if more than one sense exists for a given word, the average score across all senses is used.

Negation detection is based on the identification of explicit tokens, and is close to the *NegEx* algorithm (Chapman et al. 2001): we prepared a word list of explicit negation words based on the list from (Councill, McDonald, and Velikovich 2010), including common misspellings. A second list of pseudo-negation expressions was derived from the original NegEx list and later extended with additional examples observed from experimentation. Negation scope is determined by a punctuation mark, a known end of negation expression, or after a maximum number of tokens has been scanned (set to 5 tokens per original *NegEx* implementation). When a sentiment word is detected within the scope of a negated passage, its polarity is inverted[4].

We collected eighteen datasets covering different customer review domains from previous works in the literature: we use the benchmark IMDB film reviews dataset from (Pang, Lee, and Vaithyanathan 2002), Tripadvisor hotel reviews from (Baccianella, Esuli, and Sebastiani 2009), and Amazon.com customer reviews from different product categories collected from (Jindal and Liu 2008) and (McAuley and Leskovec 2013). Each domain was prepared with an equal number of positive and negative reviews totalling 55418 documents across all datasets. Key characteristics are given in Table 2.

In Table 3 we present accuracy results using the lexicon-based classification algorithm with each of the above lexicons. The experiment was then repeated with negation detection enabled and improvements are seen on all but one lexicon and domain (GI on Pet domain). The improvements with negation detection enabled are statistically significant

[1] www.wjh.harvard.edu/ inquirer/homecat.htm

[2] Extracted from http://sentiwordnet.isti.cnr.it/

[3] http://nlp.stanford.edu/software/tagger.shtml

[4] Source available in: https://github.com/bohana/sentlex

Table 2: Customer review datasets - all datasets contain equal number of positive and negative documents.

		Average stats per document				
Domain	Docs per class	Sent. size	Sentences	Tokens	Unique words	Words
Apparel	283	18.2	6.8	123.6	75.0	108.2
Books	1016	20.9	11.7	243.6	132.6	215.1
Electronics	1035	19.1	11.3	215.9	116.5	189.6
Music	2948	21.2	10.4	220.4	122.5	188.3
Health Products	998	17.3	6.4	110.0	67.8	98.3
Films	999	21.3	35.8	762.2	334.3	660.6
Network Equip.	998	18.5	8.2	151.7	88.2	135.0
Pet Products	998	18.2	6.3	115.2	70.9	103.3
Software	998	18.7	7.8	145.7	85.6	130.1
Hotels	1437	22.2	9.8	218.7	119.3	195.4
Car Products	2000	18.4	4.3	79.6	50.4	71.6
Baby Products	2000	19.5	5.7	110.8	65.1	99.8
DIY	2000	19.4	5.3	102.8	61.2	92.7
Jewellery	2000	16.6	3.9	65.4	42.5	58.5
Fine Foods	1999	16.7	4.4	73.6	48.1	65.6
Office	2000	19.3	5.2	100.3	60.0	89.9
Patio Furniture	2000	18.7	5.2	97.5	58.9	87.6
Toys	2000	18.6	5.1	95.8	58.5	85.9

using the Wilcoxon signed rank test ($p = 0.01$) for all lexicons tested.

Table 3: Accuracies for baseline experiment, with and without negation detection.

	Baseline			with Negation Detection		
Dataset	GI	MPQA	SWN3	GI	MPQA	SWN3
Apparel	67.49	66.61	67.49	68.55	**69.08**	**69.08**
Books	61.12	63.19	62.06	63.98	**66.49**	64.76
Electronics	64.25	67.15	63.43	68.26	**70.39**	67.97
Music	62.64	62.31	62.28	64.01	**64.21**	64.06
Health	62.73	63.13	61.02	**64.58**	66.43	64.13
Films	68.43	69.93	64.63	69.98	**71.29**	66.33
Network	64.20	66.35	60.79	66.50	**69.05**	63.24
Pet	60.80	62.81	60.75	60.50	**64.76**	62.11
Software	61.49	64.50	61.79	65.00	**68.30**	65.55
Hotels	66.04	65.03	66.98	**71.26**	71.09	71.16
Car	63.23	64.12	61.12	66.70	68.03	**68.05**
Baby	64.66	64.08	63.48	**68.63**	68.53	68.08
DIY	64.67	65.53	63.73	67.30	**68.60**	67.35
Fine Foods	62.16	63.41	65.89	65.22	67.29	**70.20**
Jewellery	67.35	67.50	69.00	70.25	71.40	**73.90**
Office	65.15	67.00	65.47	68.17	**70.83**	69.47
Patio	64.62	65.47	63.27	67.34	**69.12**	67.67
Toys	65.91	67.08	64.70	70.02	**71.39**	68.79

Next we evaluate the effects of negation as a source of negative polarity. Our approach is to boost the aggregate negative sentiment score by counting each negated passage as a negative sentiment token with a fixed score set arbitrarily at 0.1. As before, the polarity of sentiment words found within a negated window is also inverted. Results of this approach are presented in Table 4 (*Fixed* column): adding a fixed adjustment yielded performance improvements over basic negation detection on all domains. For conciseness, we show only results for the MPQA lexicon, which performed better on most domains on the baseline experiments. However similar improvements were obtained on GI and SWN3.

We investigated the possibility of setting the negation adjustment score dynamically instead of fixing this value a-priori, assuming we avail of out-of-domain data. For each of the 18 domains available, we perform a grid search on possible score values using classifier accuracy as the optimization criteria on a subset of documents from the remaining $N - 1 = 17$ domains. In Table 4 (*Grid Search* column) we show accuracies using grid search on 20 equally spaced points in the $[0 - 1]$ interval and 100 documents per class per out-of-domain dataset (3400 documents in total). This approach improved results further on 16 of the 18 domains, with one performance reduction (film reviews) and one tie (networking), which were statistically significant using the Friedman test for multiple datasets and the post-hoc Nyemeni test at $p = 0.05$ (Demšar 2006).

Table 4: Classifier accuracies - adjusting negativity of negating words (using the MPQA lexicon).

	Baselines		Negation Adjusted	
Dataset	No Negation	Negation Detection	Fixed ($score = 0.1$)	Grid Search
Apparel	66.61	69.08	72.97	**77.03**
Books	63.19	66.49	69.39	**72.54**
Electronics	67.15	70.39	74.78	**75.94**
Music	62.31	64.21	67.89	**71.10**
Health	63.13	66.43	70.79	**71.49**
Films	69.93	71.29	**71.99**	71.79
Network	66.35	69.05	**71.46**	**71.46**
Pet	62.81	64.76	68.27	**69.77**
Software	64.50	68.30	71.71	**74.41**
Hotels	65.03	71.09	75.96	**83.26**
Auto	64.12	68.03	75.00	**77.50**
Baby	64.08	68.53	73.36	**75.61**
DIY	65.53	68.60	73.00	**74.95**
Fine foods	63.41	67.29	72.65	**74.15**
Jewellery	67.50	71.40	78.10	**81.42**
Office	67.00	70.83	75.90	**77.90**
Patio	65.47	69.12	74.54	**76.09**
Toys	67.08	71.39	76.38	**79.38**

The adjustment values found via parameter search were 0.63 or 0.74, depending on the domain left out. Increasing the number of searched points in the parameter space did not yield further improvements, as illustrated in the training set accuracies for a sample domain in Figure 1 (the *x* axis is clipped to highlight the trend). Table 5 shows discovered values at search intervals of increasing granularity for a fixed training set size of 100 documents per class per $N - 1 = 17$ training domains, along with the corresponding number of domains they were found in.

When varying training size, the performance stabilizes after a relatively small amount of documents used in training: no improvements were obtained when increasing the training set size to more than 50 documents per class. Figure 2 indicates performance trends with varying training sizes and granularity.

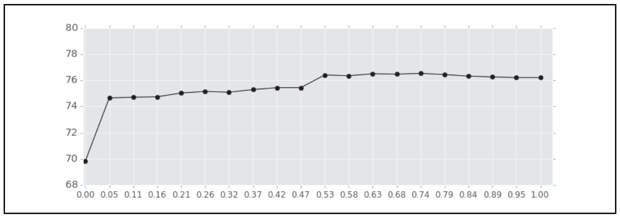

Figure 1: Training set accuracies when searching adjustment factor (Patio as the test domain).

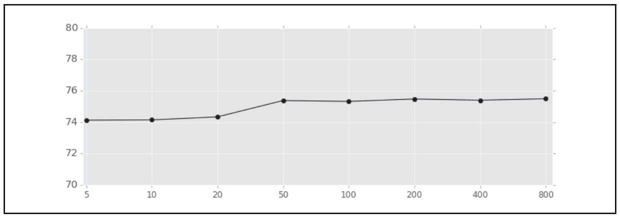

Figure 2: Mean performance by training size across all domains.

Polarity Inversion and Attenuation

As discussed in Section 2, previous experiments found better classification performance could be obtained by shifting the polarity scores of negated words, instead of inverting their value, while other authors found it beneficial to give a different treatment to positive and negative words when these are negated. In the next experiment we included a per-class *polarity multiplier* on negated words adj_{pos}, adj_{neg} with values each ranging from $[-1.0, 1.0]$. When a word is found inside a negated scope, its sentiment scores are adjusted by $Pos \cdot adj_{pos}$ and $Neg \cdot adj_{neg}$ respectively, thus allowing for a score inversion to occur when the multiplier is negative, or attenuating its value when positive.

As before, we used a grid search on the above multiplier parameters to chose values that maximised accuracy on out-of-domain data. Results in Table 6 compare this experiment (A) with our earlier baseline (with negation), and negation adjustment via grid search (C). When searching for adjustment values between $[-1.0, 1.0]$, grid search found the same pair as the optimal solution in every domain: $(adj_{pos}, adj_{neg}) = (-1.0, 0.11)$. This reflects previous results reporting improvements when negative word polarity is attenuated when negated, while positive words have their polarity inverted (Kiritchenko, Zhu, and Mohammad 2014). However, performance remained below that of (C) in all but the Films domain.

Lastly, we executed parameter search on the 3 dimensions being considered: negation adjustment and the polarity attenuation/inversion multiplier with results shown in column (B). When compared to the negation adjustment-only version from our earlier experiment (C), results were comparable but not statistically significant from it. In that regard, the approach proposed in this study (C) provides a simpler strategy in that it performs better than such adjustments used on their own (B), and while reducing the number of input variables to be considered.

Table 5: Discovered values by search space granularity.

Granularity (data points)	Discovered Score (domains)
2	1.0 (18)
5	0.75 (18)
10	0.66 (17), 0.77 (1)
20	0.63 (13), 0.74 (5)
50	0.61 (17), 0.73(1)
100	0.60 (17), 0.71 (1)

Table 6: Attenuation and Inversion.

Dataset	Baseline (with negation)	(A) Atten. Adj.	(B) Atten. + Neg Adj.	(C) Neg Adj.
Apparel	69.08	71.20	76.50	**77.03**
Books	66.49	67.37	72.34	**72.54**
Electronics	70.39	72.56	**76.62**	75.94
Music	64.21	66.15	70.96	**71.10**
Health	66.43	67.59	71.39	**71.49**
Films	71.29	72.19	**72.84**	71.79
Network	69.05	70.11	**72.91**	71.46
Pet	64.76	66.57	**70.03**	69.77
Software	68.30	69.25	74.16	**74.41**
Hotel	71.09	73.70	82.05	**83.26**
Car	68.03	69.50	76.47	**77.50**
Baby	68.53	69.71	74.89	**75.61**
DIY	68.60	69.67	**75.65**	74.95
Fine foods	67.29	68.52	**74.27**	74.15
Jewellery	71.40	73.10	80.53	**81.42**
Office	70.83	72.28	77.83	**77.90**
Patio	69.12	70.57	**76.32**	76.09
Toys	71.39	72.97	78.83	**79.38**

5 Conclusion

We have conducted an experiment on the effects of negation on lexicon-based sentiment classification of documents. We propose a novel adjustment factor based on negation occurrences as a proxy of negative sentiment polarity, and saw statistically significant performance improvements on all domains tested, by as much as 12 percentage points. Furthermore, using a parameter search on out-of-domain data proved a viable option for dynamically calibrating this adjustment, yielding significant improvements over a fixing the adjustment value a-priori on 16 out of 18 domains tested, using a relatively small training set.

Calibrating parameters using out-of-domain data may be extensible to other aspects of lexicon-based classifiers, potentially making such methods more competitive. Further refinements to negation and negation scope detection could also be beneficial for our method. Finally, our results indicate that, in the realm of customer reviews, the score adjustment for negated words has generalised well across domains. We are interested in experimenting with other types of documents and verifying if similar performance gains can be obtained on a broader section of opinionated discourse.

References

Abbasi, A.; Chen, H.; and Salem, A. 2008. Sentiment analysis in multiple languages: Feature selection for opinion classification in Web forums. *ACM Transactions on Information Systems (TOIS)* 26(3):12.

Baccianella, S.; Esuli, A.; and Sebastiani, F. 2009. Multi-facet rating of product reviews. *Advances in Information Retrieval* 461–472.

Baccianella, S.; Esuli, A.; and Sebastiani, F. 2010. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *LREC*, volume 10, 2200–2204.

Chapman, W.; Bridewell, W.; Hanbury, P.; Cooper, G.; and Buchanan, B. 2001. A simple algorithm for identifying negated findings and diseases in discharge summaries. *Journal of biomedical informatics* 34(5):301–310.

Councill, I. G.; McDonald, R.; and Velikovich, L. 2010. What's great and what's not: learning to classify the scope of negation for improved sentiment analysis. In *Proceedings of the workshop on negation and speculation in natural language processing*, 51–59. Association for Computational Linguistics.

Cui, H.; Mittal, V.; and Datar, M. 2006. Comparative experiments on sentiment classification for online product reviews. In *Proceedings of the National Conference on Artificial Intelligence*, volume 21, 1265. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999.

Dave, K.; Lawrence, S.; and Pennock, D. M. 2003. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In *Proceedings of the 12th WWW (2003)*, 528. ACM.

Demšar, J. 2006. Statistical comparisons of classifiers over multiple data sets. *The Journal of Machine Learning Research* 7:30.

Denecke, K. 2009. Are SentiWordNet scores suited for multi-domain sentiment classification? In *ICDIM 2009.*, 1–6. IEEE.

Ding, X.; Liu, B.; and Yu, P. S. 2008. A holistic lexicon-based approach to opinion mining. In *Proceedings of the international conference on Web search and web data mining*, 231–240. ACM.

Gezici, G.; Yanikoglu, B.; Tapucu, D.; and Saygın, Y. 2012. New features for sentiment analysis: Do sentences matter? In *SDAD 2012 The 1st International Workshop on Sentiment Discovery from Affective Data*, 5.

Goyal, A., and Daumé III, H. 2011. Generating semantic orientation lexicon using large data and thesaurus. In *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis*, 37–43. Association for Computational Linguistics.

Hatzivassiloglou, V., and McKeown, K. R. 1997. Predicting the semantic orientation of adjectives. In *35th Annual Meeting of the ACL*, ACL '98, 174–181. Association for Computational Linguistics.

Jindal, N., and Liu, B. 2008. Opinion spam and analysis. In *Procs of the conference on Web search and Web data mining (WSDM'08)*, 219–230. ACM.

Kennedy, A., and Inkpen, D. 2006. Sentiment classification of movie reviews using contextual valence shifters. *Computational Intelligence* 22(2):110–125.

Kiritchenko, S.; Zhu, X.; and Mohammad, S. M. 2014. Sentiment analysis of short informal texts. *Journal of Artificial Intelligence Research* 723–762.

Liu, B., and Zhang, L. 2012. A survey of opinion mining and sentiment analysis. In *Mining text data*. Springer. 415–463.

Martin, J. R., and White, P. R. 2005. *The language of evaluation*. Palgrave Macmillan Basingstoke and New York.

McAuley, J., and Leskovec, J. 2013. Hidden factors and hidden topics: understanding rating dimensions with review text. In *Proceedings of the 7th ACM conference on Recommender systems*, 165–172. ACM.

Mohammad, S. M., and Turney, P. D. 2013. Crowdsourcing a word–emotion association lexicon. *Computational Intelligence* 29(3):436–465.

Morante, R., and Sporleder, C. 2012. Modality and negation: An introduction to the special issue. *Computational linguistics* 38(2):223–260.

Paltoglou, G., and Thelwall, M. 2010. A study of information retrieval weighting schemes for sentiment analysis. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, 1386–1395. ACL.

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

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

Pang, B.; Lee, L.; and Vaithyanathan, S. 2002. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of 2002 EMNLP*, 79–86. ACL.

Poria, S.; Cambria, E.; Winterstein, G.; and Huang, G.-B. 2014. Sentic patterns: Dependency-based rules for concept-level sentiment analysis. *Knowledge-Based Systems* 69:45–63.

Poria, S.; Cambria, E.; and Gelbukh, A. 2015. Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis. In *Proceedings of EMNLP*, 2539–2544.

Potts, C. 2011. On the negativity of negation. In *Proceedings of SALT*, volume 20, 636–659.

Stone, P. J.; Dunphy, D. C.; Smith, M. S.; Ogilvie, D. M.; and Others. 1966. *The general inquirer: A computer approach to content analysis*. MIT Press, Cambridge, MA.

Taboada, M.; Brooke, J.; Tofiloski, M.; Voll, K.; and Stede, M. 2011. Lexicon-based methods for sentiment analysis. *Computational linguistics* 37(2):267–307.

Whitelaw, C.; Garg, N.; and Argamon, S. 2005. Using appraisal groups for sentiment analysis. In *Proceedings of the 14th ACM international conference on Information and knowledge management*, 625–631. ACM.

Wiegand, M.; Balahur, A.; Roth, B.; Klakow, D.; and Montoyo, A. 2010. A survey on the role of negation in sentiment analysis. In *Proceedings of the workshop on negation and speculation in natural language processing*, 60–68. ACL.

Wilson, T.; Wiebe, J.; and Hoffmann, P. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Procs of HLT/EMNLP'05*, 354. ACL.