2011

Domain Independent Sentiment Classification with Many Lexicons

Bruno Ohana
Technological University Dublin

Brendan Tierney
Technological University Dublin

Sarah Jane Delany
Technological University Dublin, sarahjane.delany@tudublin.ie

Recommended Citation

B. Ohana, B. Tierney, and S. J. Delany. (2011) Domain independent sentiment classification with many lexicons. In 4th International Symposium on Mining and Web at 25th International Conference on Advanced Information Networking and Applications (AINA), pages 632–637. IEEE Computer Society. doi:10.1109/WAINA.2011.103

Domain Independent Sentiment Classification with Many Lexicons

Bruno Ohana, Brendan Tierney and Sarah-Jane Delany
School of Computing
Dublin Institute of Technology
Dublin, Ireland
bruno.ohana@student.dit.ie; {brendan.tierney, sarahjane.delany}@dit.ie

***Abstract*— Sentiment lexicons are language resources widely used in opinion mining and important tools in unsupervised sentiment classification. We present a comparative study of sentiment classification of reviews on six different domains using sentiment lexicons from different sources. Our results highlight the tendency of a lexicon's performance to be imbalanced towards one class, and indicate lexicon accuracy varies with the target domain. We propose an approach that combines information from different lexicons to make classification decisions and achieve more robust results that consistently improve our baseline across all domains tested. These are further refined by a domain independent score adjustment that mitigates the effect of the recall imbalance seen on some of the results.**

Keywords: Opinion Mining, Sentiment Lexicon, Sentiment Classification, Natural Language Processing, Multiple Classifier Systems

I. INTRODUCTION

A sentiment lexicon is a database that associates words and expressions with information on their evaluative capacity and positive or negative orientation, with applications in a number of opinion mining tasks such as sentiment extraction, sentiment classification and subjectivity detection. Approaches to building lexicons proposed in the literature range from the manual annotation of word lists to automated techniques leveraging an existing language resource such as document corpora or thesauri. Because opinion lexicons embed prior knowledge about the sentiment of a term or expression, they are particularly useful in cases when no training data is available. It is thus an important component of unsupervised sentiment classification methods.

In this research we investigate the role of sentiment lexicons when applied to sentiment classification of user generated reviews from different domains. Our contributions to research comprise of a comparative study of lexicon based sentiment classification on multiple domains showing that classification performance varies with the chosen lexicon and the domain it is applied to, and that lexicons show a tendency to perform better on either positive or negative documents while underperforming on the other category. In addition, from this observation we propose an approach that takes into account predictions from different lexicons by combining them in a classifier ensemble, and we further extend it by introducing a score adjustment factor based on a term's relative frequency of occurrence extracted from a corpus. We obtain improvements over the baseline results across all domains, suggesting that leveraging information from many lexicons is a more robust method for sentiment classification when applied to yet unseen domains.

In the next section we discuss related work in the literature of opinion lexicons and their applications to cross domain techniques for sentiment classification. We then present our experiment setup and the results of a baseline classification task using a selection of lexicons available from the literature, plus an additional lexicon built for this research. We introduce our new proposed approaches and discuss our findings and avenues for future work.

II. RELATED WORK

A. Sentiment Lexicons

The *semantic orientation* of a term [21] indicates its capacity for carrying positive or negative evaluative value. It is possible for this information to exist a priori with relative independence from the context it may appear, as seen on words such as "excellent" or "terrible". For this reason knowledge of such terms can be a useful when identifying sentiment and is a motivation for the development of collections of opinionated terms into a sentiment lexicon.

Sentiment lexicons exist as manually annotated databases such as the General Enquirer [17] mapping terms in the English language into semantic categories, including sentiment orientation. Initially compiled to assist research on social studies, it has proven useful on opinion mining research and is regarded as a highly accurate lexicon used as a baseline for comparisons [1][31]. Other ad-hoc manual resources were generated for specific research [6]. However the collection and annotation of a large sized lexicon is an expensive and time consuming task and has motivated research in automated methods that leverage existing language resources to build or expand existing lexicons. Early work seen in [21] proposes the extension of a choice of seed words by evaluating the presence of connecting terms ("and", "or", "but") between adjectives in a large document corpus. Similar corpus based methods exploring other linguistic patterns were proposed in [8], and in [22] an extension of this approach suggests using a supervised learning technique to automatically identify language expressions that correlate terms, and then extract terms based on such patterns. Term proximity is also investigated in [10] where a list of seed terms is extended according to a measure of co-occurrence with other terms in a document corpus. A

lexicon based on proximity measures using documents obtained from search engine results is proposed in [16] and [14].

Other approaches explore semantic relations in existing language resources, with the WordNet database [9] being a prominent one: by traversing WordNet's term relationships such as synonyms and antonyms, it is possible to extend a lexicon from a list of seed terms. This approach is seen in studies employing WordNet-based lexicons for specific opinion mining tasks [25], [24].

As observed in [23], Wordnet's semantic relationships form a highly disconnected graph, thus imposing limitations to extending a lexicon based on this information alone. An approach to overcome this issue is proposed in the SentiWordNet lexicon [3], where descriptive text contained in term glosses is used to train a committee of supervised learning classifiers and predict the orientation of not yet seen terms.

Methods that use a thesaurus for building sentiment lexicons are also found in the literature: In [31] the *Macquaire* thesaurus is used to extend a list of seed words obtaining a high precision, high coverage lexicon, while [32] use the *Roget* thesaurus to extend a list of words representing different emotion categories.

B. Sentiment Classification

The objective of sentiment classification is to predict the overall sentiment orientation conveyed in a piece of text such as a user review, blog post or editorial. Several supervised learning approaches were proposed in the past decade with considerable success: early work from Pang et al. [6] presents a series of experiments evaluating various supervised learning algorithms for classifying film reviews as positive or negative. Work from [26] shows that higher order n-gram vectors can obtain good results when significantly larger data sets are available for training. A similar approach is seen in [27]. The addition of other features derived from parts of speech to a supervised learning model is explored in [10]. In [18] a method that detects and scores patterns in part of speech is applied to derive features for sentiment classification, with a similar idea applied to opinion extraction for product features seen in [28]. The work of [4] and [2] present experiments using similar techniques and improve their results by adding a feature selection step to the classifier training stage.

However, experimental results seen in [2] show that supervised learning techniques using in-domain data do not scale well across different domains: words that make good predictors within a domain are not easily generalized, for example in the case of actor or director names being good predictors of author opinion on film reviews and thus making useful features to train a classifier but which naturally will have limited applicability on an unrelated domain. In this context, interest on techniques that rely less on domain knowledge has grown considerably. These include methods that leverage properties of natural language, discourse analysis and lexicons.

The use of lexicons in sentiment classification is seen on an early experiment reported in [6] using term counting to predict the sentiment in movie reviews. Similarly [13] present several results on applying lexicon-based approaches to sentiment classification, and the authors also demonstrate how lexicon based and supervised learning can be combined as different sources of information to obtain better classification results. A similar approach is seen in [1] applied to cross domain sentiment classification: here a WordNet-based lexicon is applied in conjunction with supervised learning methods to produce an ensemble of classifiers for document and sentence level sentiment classification on different domains and genres. In [5] different scoring techniques are evaluated on the domain of film reviews using the SentiWordNet lexicon.

Work closely related to our research is seen in [14], where a sentiment classification experiment across different domains uses a custom built sentiment lexicon while also exploring the use of linguistic clues such as negation and valence shifting terms.

III. EXPERIMENT

Our research aims at establishing how lexicons built using different methods and knowledge sources perform on a sentiment classification experiment across many domains. Our choice of lexicons is a mixture of building techniques available in the literature, from manually compiled resources to automated build methods. We use the *General Inquirer* [17] and the *Subjectivity Clues* [29] lexicons. The later is a collection of opinion bearing terms gathered from manually annotated resources and extended via automatic extraction from text and thesauri. *SentiWordNet* [3] is based on the WordNet database and uses its semantic relations to expand a list of seed words and further expanded by examining a term's textual explanation (glosses) present in the database. Finally we introduce a sentiment lexicon based on the Moby public domain thesaurus for the English language (*http://icon.shef.ac.uk/Moby*). The lexicon is built by exploring the grouping of semantically related words available in Moby to extend a list of core words. It is included in the experiment as a means to assess how sentiment classification performance varies with the build method and the underlying language resource it is built upon, and thus the focus of our discussion will be on contrasting its results on sentiment classification with that of other lexicons rather than the lexicon's term accuracy to a gold standard. The lists of opinionated terms are initialized with 56 positive and 55 negative words manually annotated by the authors covering Lemke's semantic categories for evaluation presented in [11]. The lists are then expanded by adding related terms from Moby while removing terms that appear on both categories. After one iteration of the expansion we obtain the lexicon presented here.

Table 1 presents comparative figures from each lexicon. All lexicons indicate opinion polarity by means of a real valued positive and negative score ranging from 0 to 1. For lexicons where no numeric score is provided a value of 1 is assigned to the category the term belongs (positive or negative) and zero otherwise. Agreement is calculated on the basis of whether a term belongs to positive, negative or neutral classes on the two lexicons – i.e. the values for

opinion scores may be different, so long as the indicative sentiment is the same (a positive score of 0.5 on lexicon A and 0.7 on lexicon B would be considered in agreement). Where both positive and negative scores are present for the same term (indicating sentiment for different senses), the highest score is considered for agreement.

TABLE I. COMPARISON OF LEXICONS IN EXPERIMENT

Lexicon	Adjectives					Verbs				
	Pos	Neg	Total	∩ GI	Agreement (%)	Pos	Neg	Total	∩ GI	Agreement (%)
GI [17]	771	800	2514	N/A	N/A	406	702	2331	N/A	N/A
Subj. Clues [29]	1533	2513	3944	1432	83.51	742	1541	2272	831	85.92
SWN [3]	7668	9660	21436	2058	55.78	2146	2623	11306	2147	49.97
Moby	3032	3963	6966	1354	53.69	500	614	1114	414	31.4

Using the General Enquirer (GI) as a baseline lexicon for our comparisons, we see that the rate of agreement and the number of terms in common (see ∩ GI column) vary considerably for lexicons built using different methods and based on different knowledge sources. GI and the Subjectivity Clues lexicon are close in size and agreement as the later incorporates all of GI data. SentiWordNet (SWN) and Moby are built upon different knowledge resources and show a higher rate of disagreement with GI. Such differences can be attributed in part to inaccuracies in the build method and the underlying knowledge resource itself. However some authors support the view that opinion polarity for certain terms can be ambiguous, and that is seen on high levels of inter annotator disagreement [1], thus it is also possible to attribute the disagreement to different knowledge sources taking different viewpoints on the predominant sentiment orientation of a term.

A. *Sentiment Classification by Term Scoring*

We determine document sentiment based in the sum of the scores for terms found to carry positive or negative orientation under the assumption that author sentiment is correlated to the choice and number of opinionated terms present in the text. Similar scoring approaches are seen on previous research in [5][6][12][13]. The scoring method extracts opinion scores from a lexicon for each matching term in the document, and the class with the highest aggregated total score determines overall sentiment.

Information from the sentiment lexicons is related mostly to individual terms, and categorized by part of speech such as adjective, verb or noun. To compute this information from plain text we employ an automatic part of speech tagger application: the Stanford POS Tagger (*http://nlp.stanford.edu/software/tagger.shtml*). We have chosen to include adjectives and verbs in the sentiment classification scoring, as these have been seen to be good indicators of opinion in documents [6][14].

Negation detection is also an important factor when predicting sentiment from term information. Clearly, the sentences “this book is great” and “this book is not so great” convey very different sentiment despite containing the same positive term. We employed a variation of the *NegEx* algorithm [7] for detecting sections of the document where sentiment is being affected by a negating expression. NegEx works by scanning the text for a collection of known negating expressions, and marking terms as being negated according to a “window” that determines how many terms ahead or backwards are affected. When a negated term is found the affected terms have their opinion scores inverted.

B. *Data Sets*

The experiment is executed on six data sets containing user generated reviews from different domains: the movie review data set extracted from IMDB [6]; the TripAdvisor data set of hotel reviews presented in [19] and four additional data sets of reviews in books, apparel, music and consumer electronics domains extracted from the data presented in [15]. For these data sets, users could score their opinions on a scale of 1 to 5, only reviews rating 4 and 5 were chosen for the positive class while reviews scoring 1 and 2 were added to the negative class. The proportion of positive and negative reviews on all data sets is 50% each; the complete size of each data set is: music: 5902; apparel: 566; film: 2000; hotel: 2874; electronics: 2072; books: 2034.

C. *Baseline Results*

In Table 2 we detail the baseline accuracy results obtained from using term scoring on each of the four lexicons discussed previously. The best result on each domain is highlighted.

We observe first that using the same scoring method no single lexicon performs best across all domains, for example SWN wins in the electronics, hotels and music domains while Subjectivity Clues wins on books, films and apparel. This suggests a specific match of opinion terms more likely to appear in a domain and their opinion information from a given lexicon are contributing to the final accuracy results. Secondly, an imbalance on class recall can be noticed on many of the results. For instance, GI and Subjectivity Clues lexicons show high recall on the positive class but very poor results on the negative class on all domains; SWN shows the same trend but results are more balanced on the film domain while Moby shows an inverted trend. Performance imbalance results were also seen in [12] on a single lexicon experiment, where the scoring results for cross domain sentiment classification were adjusted by applying a constant multiplier.

Our results however show the class imbalance is not constant across different domains, and is dependent on the choice of lexicon. Thus applying a constant correction factor to the resulting score may be limited in its benefits as it may not guarantee good results on a yet unseen domain or when using a new lexicon.

TABLE II. BASELINE SENTIMENT CLASSIFICATION RESULTS

Data Set	Lexicon	Accuracy (%)	Recall Pos. Class	Recall Neg. Class
Films	GI	66.85	87.70	46.00
	Subj. Clues	**68.2**	83.00	53.40
	SWN	65.65	64.90	66.40
	Moby	58.95	7.90	96.00
Hotels	GI	65.97	99.30	32.64
	Subj. Clues	67.15	99.58	34.73
	SWN	**71.68**	96.31	47.04
	Moby	65.66	48.64	82.67
Electronics	GI	63.85	91.99	35.71
	Subj. Clues	66.8	93.63	39.96
	SWN	**67.18**	76.93	57.43
	Moby	53.96	45.73	62.55
Books	GI	60.52	87.41	33.63
	Subj. Clues	**63.72**	88.79	38.64
	SWN	62.05	70.01	54.08
	Moby	57.82	55.95	59.69
Apparel	GI	64.31	90.11	38.52
	Subj. Clues	**65.55**	95.05	36.04
	SWN	64.13	74.56	53.71
	Moby	54.24	45.94	62.54
Music	GI	60.74	92.38	29.11
	Subj. Clues	61.71	94.17	29.24
	SWN	**65.08**	81.67	48.49
	Moby	59.64	56.22	63.03

D. *Combining Lexicons*

Based on our previous observations we argue that each lexicon, with its particular choice of terms and encoded sentiment information is uniquely capable of reaching a classification decision. Such decisions can be compared against that of other lexicons in a voting scheme typical of classifier ensembles. As pointed out in [20], this approach could yield more robust results if classifiers based on different lexicons are independent in how they produce classification errors. We choose three distinct lexicons based on their differing build methods and rate of disagreement with the General Inquirer lexicon as shown in Table 1: SentiWordNet, Subjectivity Clues and Moby. The rationale is that this selection will maximize the use of unique information contained in each lexicon when making a prediction.

The document score calculation obtained from each lexicon provides normalized real valued positive and negative scores which allow us to experiment on different approaches to majority voting. We obtain a prediction according to three of the schemes presented in [33]: on *majority voting* each prediction receives an unweighted vote and the class with highest votes is selected; the *sum rule* states that the class with the highest aggregated score (obtained from the document scoring using each lexicon) is selected; while the *max rule* chooses the class whose score is the highest obtained from the scores of each individual lexicon.

The above ensemble schemes assume the use of classifiers that produce posterior class probabilities. Our experiment uses normalized scores and while it is possible to transform those into an estimation of posterior probabilities (see [34] for a survey of techniques) they would require availing of a training data set on a given domain, which may not give us the desired estimates on a cross domain scenario. For this experiment we treat the calculated normalized scores as uncalibrated posterior probabilities. We give the results from each approach in Table 3 and compare them against the best baseline obtained on each domain.

TABLE III. ACCURACY COMPARISON – COMBINING LEXICONS

	Film	Hotel	Elect.	Books	Apparel	Music
Maj. Vote	68.55	73.5	68	63.77	66.25	65.62
Sum Rule	**69.6**	**80.23**	**69.35**	**65.63**	**68.37**	**67.55**
Max Rule	67.8	79.82	62.36	62.88	63.07	65.38
Best Baseline	68.2	71.68	67.18	63.72	65.55	65.08

When using the sum rule, classification accuracies improved over the best baseline on all domains.

E. *Adjusting Scores Based on Term Frequency*

One possible reason for imbalance on scores is the fact that certain terms do naturally occur more often in language than others, irrespective of what overall sentiment a given text is conveying. Such terms can negatively affect classification accuracy should they appear in lexicon as a non neutral term.

One approach to test this hypothesis is to adjust the scores of a term according to how frequently they occur on language regardless of the opinion of underlying text. We estimate term frequency by calculating relative frequencies of all terms in a lexicon based on frequency data extracted from the Brown document corpus for the news, reviews and editorial categories, under the assumption the data is representative from a mixture of natural language text from different domains likely to appear on opinionated text. In our adjustment terms more likely to appear in arbitrary text have their scores reduced according to the formula given in (1):

$$s_{adj}(w) = s(w) * (1 - \sqrt{freq(w)}) \quad (1)$$

Where *s(w)* is the unadjusted score obtained from a lexicon and *freq(w)* is the frequency of word *w* relative to that of the most frequent term found in the lexicon, computed from corpus data, valued between 0 and 1. Thus the impact of a highly frequent term to the overall document score is reduced according to how frequently it is expected to occur. We present the results of frequency adjusted scores for the majority voting schemes in Table 4.

TABLE IV. ACCURACY COMPARISON – FREQUENCY ADJUSTMENT

	Film	Hotel	Elect.	Books	Apparel	Music
Maj. Voting + Freq	69.6	76.62	**70.08**	65.39	67.49	67.6
Sum Rule + Freq	**69.9**	**80.79**	68.68	**66.22**	**68.73**	**68.32**
Max Rule + Freq	66.2	80.27	61.34	63.27	63.6	66.18
Sum Rule	69.6	80.23	69.35	65.63	68.37	67.55
Best Baseline	66.85	71.68	67.18	63.72	65.55	65.08

IV. DISCUSSION

A. Class Imbalance

Results on Table 2 show that recall for some lexicons' predictions display a considerable class imbalance, not seen on accuracy figures alone. This behavior is unwanted on practical application where misclassification costs vary with class, or class distribution may be skewed. To measure the effects of our methods on class imbalance, we consider the minimum class recall across the different methods tested. Table 5 compares the minimum recall obtained on either positive or negative class obtained from the experiment with best accuracy results from each approach investigated.

TABLE V. MIN. CLASS RECALL FOR EACH EXPERIMENT

	Film	Hotel	Electr.	Books	Apparel	Music
Sum Rule + Freq	61.1	**66.39**	**62.26**	**60.08**	**60.78**	**54.29**
Sum Rule	**65.6**	64.37	59.36	55.75	57.95	50.66
Best Baseline	53.4	47.04	57.43	38.64	36.04	48.49

The results show a reduction on the worst case recall obtained by a single class when combining lexicons, and even though gains in classification accuracy seen on frequency adjustment are small, they still provide improvements over the worst case class recall on all but one domain, making correct predictions more evenly distributed across positive and negative documents.

B. Statistical Comparison of Results

Our experiment generates performance results across different data sets, and to measure the statistical validity of the improvements obtained we use the *Wilcoxon Signed-Rank* test based on the ranking of differences of paired results of a performance metric across different tests. It is argued in [30] that it is a more suitable test for experiments on different data sets. We first rank the difference on each classifier's results as illustrated on Table 6.

TABLE VI. WILCOXON SIGNED-RANK TEST

	Film	Hotel	Electr.	Books	Apparel	Music
Sum Rule	**69.6**	**80.23**	**69.35**	**65.63**	**68.37**	**67.55**
Best Baseline	66.85	71.68	67.18	63.72	65.55	65.08
Difference (÷100)	0.0275	0.0855	0.0217	0.0191	0.0282	0.0247
Rank	5	6	2	1	4	3

For each classifier, we sum the ranks for all cases when the classifier outperforms the other and calculate the Wilcoxon statistic *W*, which is the smallest of the ranked sums. This can be compared to the critical value for N=6 data sets. Results for the *sum rule* outperform the best baseline in every test, thus $W = 0$ and is below the critical value. The null hypothesis of no reliable difference between results can be rejected and the improvement is considered significant with a confidence level of α=0.05.

Moreover, calculating the same statistic for the frequency adjusted *sum rule* is not significant when compared to the non-adjusted version, as the results are improved on all but one data set. We note however that the improvements on either version are significant when compared to the best single-lexicon baseline.

V. CONCLUSIONS AND FUTURE WORK

In this research we show how different lexicons perform on the task of document sentiment classification on different domains. Our results indicate that given a fixed scoring method the performance of a given lexicon is dependent on the domain it is applied to. Additionally, a lexicon's tendency to perform better on either positive or negative predictions can also depend on the domain and lexicon used.

By combining the predictions of classifiers using separate lexicons using the sum of all scores as the predictor we obtained consistently better accuracy results than any method based on a single lexicon, across the six domains in our experiment. Moreover, by introducing a score adjustment based on term frequencies computed from a separate corpus, we were able to mitigate imbalance issues on class recall making it a more promising approach for cases when misclassification costs can vary with class. This suggests using many lexicons from different knowledge sources can be a more robust approach for cross domain sentiment classification. Exploring this technique with a wider variety of lexicons and determining criteria for adding lexicons to an ensemble are interesting extensions of this research.

Mitigating some limiting factors of lexicon based approaches such as leveraging opinion present in expressions not constrained to a single word and improving the scoring algorithm to account for indicative clues of subjectivity and document structure are also strategies we would like to explore in the future. We see sentiment lexicons as a key building block of domain independent, unsupervised sentiment classification, and its effective use will contribute to better methods in this area.

ACKNOWLEDGMENT

We wish to thank Professor Bing Liu from University of Chicago Illinois and his team for providing access to a large set of user generated review data, enabling the creation of some of the data sets used in this research.

REFERENCES

[1] A. Andreevskaia and S. Bergler, "When specialists and generalists work together: Overcoming domain dependence in sentiment tagging," *Proceedings of ACL-08: HLT*, pp. 290-298, 2008.

[2] A. Aue and M. Gamon, "Customizing sentiment classifiers to new domains: a case study," in *Proceedings of RANLP*, 2005.

[3] A. Esuli and F. Sebastiani, "Sentiwordnet: A publicly available lexical resource for opinion mining.", *5th Conference on Language Resources and Evaluation (LREC'06), 2006*, pp. 417-422.

[4] Abbasi, H. Chen, and A. Salem, "Sentiment analysis in multiple languages: Feature selection for opinion classification in Web forums," *ACM Transactions on Information Systems (TOIS)*, vol. 26, no. 3, p. 12, 2008.

[5] B. Ohana and B. Tierney, "Sentiment classification of reviews using SentiWordNet," in *9th. IT & T Conference, 2009*, p. 13.

[6] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up?: sentiment classification using machine learning techniques," *EMNLP '02: Empirical methods in natural language processing. Association for Computational Linguistics*, 2002, pp. 79-86.

[7] Chapman W, Bridewell W, Hanbury P, Cooper G, Buchanan B. "Evaluation of Negation Phrases in Narrative Clinical Report". *Proceedings of 2001 AMIA Symposium*, 2001, 105-109.

[8] D. Lin and S. Zhao, "Identifying synonyms among distributionally similar words," *in Proceedings of IJCAI-03*, 2003, pp. 1492-1493.

[9] G. A. Miller, "WordNet: a lexical database for English," *Communications of the ACM*, vol. 38, no. 11, p. 41, 1995.

[10] H. Yu and V. Hatzivassiloglou, "Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences," *Proceedings of EMNLP*, vol. 3, 2003, pp. 129-136.

[11] J. L. Lemke, "Resources for attitudinal meaning: Evaluative orientations in text semantics," *Functions of language*, vol. 5, no. 1, pp. 33-56, 1998.

[12] K. Voll and M. Taboada, "Not all words are created equal: Extracting semantic orientation as a function of adjective relevance," *AI 2007: Advances in Artificial Intelligence*, 2007, pp. 337-346.

[13] Kennedy A. and Inkpen D. "Sentiment Classification of Movie Reviews Using Contextual Valence Shifters". *Computational Intelligence*, 2006, Vol. 22, 110–125.

[14] M. Taboada, K. Voll, and J. Brooke, "Extracting sentiment as a function of discourse structure and topicality", *JNLE*, 2008.

[15] N. Jindal and B. Liu, "Opinion spam and analysis," in *Proceedings of the international conference on Web search and data mining - WSDM'08*. ACM, 2008, pp. 219-230.

[16] P. D. Turney and M. L. Littman. "Un-supervised learning of semantic orientation froma hundred-billion-word corpus". *Technical Report EGB-1094, National Research Council Canada*, 2002.

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

[18] P. Turney. "Thumbs up or Thumbs down? Sentiment Orientation Applied to Unsupervised Classification of Reviews". *40th Annual Meeting of the ACL*, 2002.

[19] S. Baccianella, A. Esuli, and F. Sebastiani, "Multi-facet rating of product reviews," *Advances in Information Retrieval*, 2009, pp. 461-472.

[20] T. Dietterich, "Ensemble methods in machine learning," *Multiple classifier systems*, pp. 1-15, 2000.

[21] V. Hatzivassiloglou and K. R. McKeown, "Predicting the semantic orientation of adjectives," in *Proceedings of the eighth conference on European chapter of the ACL*, 1997, pp. 174-181.

[22] Y. Qu and G. Grefenstette, "Finding ideographic representations of Japanese names written in Latin script via language identification and corpus validation," *42nd Annual Meeting of the ACL*, 2004, p. 183.

[23] Rao, D., and Ravichandran, D. "Semi-Supervised Polarity Lexicon Induction". *In Preceedings of the 12th Conference of the European Chapter of the ACL*. 2009, Athens, Greece, p. 675-682.

[24] N. Godbole, M. Srinivasaiah, and S. Skiena, "Large-scale sentiment analysis for news and blogs," *ICWSM'07*, 2007.

[25] Kamps, J., Marx, M., Mokken, R.J., and De Rijke, M. "Using WordNet to measure semantic orientation of adjectives". *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004).* 2004, Vol 4, p. 1115-1118.

[26] Cui, H., Mittal, V., and Datar, M. "Comparative Experiments on Sentiment Classification for Online Product Reviews". *Proceedings of the National Conference on Artificial Intelligence*, 2006, AAAI Press, Vol. 21, p. 1265-1270.

[27] Dave, K., Lawrence, S., and Pennock, D. "Mining the Peanut Gallery: Opinion Extraction and Semantic Classification in Product Reviews". *Proceedings of the 12th International conference on the World Wide Web - ACM WWW2003*, 2003.

[28] Yi, J., Nasukawa, T., Bunescu, R. and Niblack, W. "Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques". *Third IEEE International Conference on Data Mining - ICDM 2003*. 2003, p.427-434.

[29] T. Wilson, J. Wiebe, and P. Hoffmann, "Recognizing Contextual Polarity in Phrase-level Sentiment Analysis", *Human Language Technology and Empirical Methods in Natural Language Processing*. ACL, 2005, p. 354.

[30] J. Demšar, "Statistical Comparisons of Classifiers over Multiple Data sets", The Journal of Machine Learning Research, vol. 7, p. 30, 2006.

[31] S. Mohammad, C. Dunne, and B. Dorr, "Generating high-coverage semantic orientation lexicons from overtly marked words and a thesaurus", *Proceedings of the 2009 EMNLP*: Volume 2. ACL, 2009, pp. 599-608.

[32] S. Aman and S. Szpakowicz, "Using Roget's thesaurus for fine-grained emotion recognition", *Third International Conference on Natural Language Processing IJCNLP*, 2008, pp. 296-302.

[33] J. Kittler, M. Hatef, R. P. W. Duin, and J. Matas, "On Combing Classifiers", *IEEE transactions on pattern analysis and machine intelligence*, vol. 20, no. 3, pp. 226-239, 1998.

[34] B. Zadrozny and C. Elkan, "Transforming Classifier Scores into Accurate Multiclass Probability Estimates", *8th SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2002, pp. 694-699.