2009-10-01

Sentiment Classification of Reviews Using SentiWordNet

Bruno Ohana
Technological University Dublin, bohana@gmail.com

Brendan Tierney
Technological University Dublin, brendan.tierney@tudublin.ie

Recommended Citation

Ohana, B. & Tierney, B. (2009) Sentiment classification of reviews using SentiWordNet. *9th. IT&T Conference,* Technological University Dublin, Dublin, Ireland, 22-23 October. doi:10.21427/D77S56

Sentiment Classification of Reviews Using SentiWordNet

Bruno Ohana [1], Brendan Tierney [2]

[1] Dublin Institute of Technology,
School of Computing
Kevin St. Dublin 8, Ireland
bohana@gmail.com

[2] Dublin Institute of Technology,
School of Computing
Kevin St. Dublin 8, Ireland
brendan.tierney@dit.ie

Abstract

Sentiment classification concerns the use of automatic methods for predicting the orientation of subjective content on text documents, with applications on a number of areas including recommender and advertising systems, customer intelligence and information retrieval. SentiWordNet is an opinion lexicon derived from the WordNet database where each term is associated with numerical scores indicating positive and negative sentiment information. This research presents the results of applying the SentiWordNet lexical resource to the problem of automatic sentiment classification of film reviews. Our approach comprises counting positive and negative term scores to determine sentiment orientation, and an improvement is presented by building a data set of relevant features using SentiWordNet as source, and applied to a machine learning classifier. We find that results obtained with SentiWordNet are in line with similar approaches using manual lexicons seen in the literature. In addition, our feature set approach yielded improvements over the baseline term counting method. The results indicate SentiWordNet could be used as an important resource for sentiment classification tasks. Additional considerations are made on possible further improvements to the method and its use in conjunction with other techniques.

Keywords: Sentiment Analysis, Opinion Mining, SentiWordNet, Data Mining, Knowledge Discovery

1 Introduction

Opinion mining research considers the computational treatment of subjective information contained in text. With the rapid growth of available subjective text on the internet in the form of product reviews, blog posts and comments in discussion forums, opinion mining can assist in a number of potential applications in areas such as search engines, recommender systems and market research.

One approach for detecting sentiment in text present in literature concerns the use of lexical resources such as a dictionary of opinionated terms. SentiWordNet [6] is one such resource, containing opinion information on terms extracted from the WordNet database and made publicly available for research purposes. SentiWordNet is built via a semi supervised method and could be a valuable resource for performing opinion mining tasks: it provides a readily available database of term sentiment information for the English language, and could be used as a replacement to the process of manually deriving ad-hoc opinion lexicons. In addition, SentiWordNet is built upon a semi automated process, and could easily be updated for future versions of WordNet, and for other languages where similar

lexicons are available. Thus, an interesting research question is to assess how effective is SentiWordNet in the task of detecting sentiment in comparison to other methods, and what are the potential advantages that could be obtained from this approach.
This paper proposes a method for applying SentiWordNet to derive a data set of document metrics and other relevant features, and performs an experiment on sentiment classification of film reviews using the polarity data set introduced in [14]. We present and discuss the results obtained in light of similar research performed using manually built lexicons, and investigate possible sources of inaccuracies with this method. Further analysis of the results revealed opportunities for improvements to this approach, which are presented in our concluding remarks.

2 Sentiment Classification

Sentiment classification is an opinion mining activity concerned with determining what, if any, is the overall sentiment orientation of the opinions contained within a given document. It is assumed in general that the document being inspected contains subjective information, such as in product reviews and feedback forms. Opinion orientation can be classified as belonging to opposing positive or negative polarities – positive or negative feedback about a product, favorable or unfavorable opinions on a topic – or ranked according to a spectrum of possible opinions, for example on film reviews with feedback ranging from one to five stars.
Supervised learning methods using different aspects of text as sources of features have been proposed in the literature. Early work seen in [13] presents several supervised learning algorithms using bag-of-words features common in text mining research, with best performance obtained using support vector machines in combination with unigrams. Classifying terms from a document into its grammatical roles, or parts of speech has also been explored: In [21] part of speech information is used as part of a feature set for performing sentiment classification on a data set of newswire articles, with similar approaches attempted in [10], [7] and [16], on different data sets. On [20] 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 [4]. Separation of subjective and objective sentences for the purposes of improving document level sentiment classification are found in [14], where considerable improvements were obtained over a baseline word vector classifier. Other studies focus on the correlation of writing style to overall sentiment, taking into account the use of colloquialisms and punctuation that may convey sentiment. In [22] a lexicon of colloquial expressions and a regular expression rule base is created to detect unique opinion terms such as unusual spellings (“greeeat”) and word combinations (“supergood”). In [1] document statistics and features measuring aspects of writing style are combined with word vectors to obtain considerable improvements over a baseline classifier on a data set of film reviews.

2.1 Opinion Lexicons

Opinion lexicons are resources that associate sentiment orientation and words. Their use in opinion mining research stems from the hypothesis that individual words can be considered as a unit of opinion information, and therefore may provide clues to document sentiment and subjectivity. Manually created opinion lexicons were applied to sentiment classification as seen in [13], where a prediction of document polarity is given by counting positive and negative terms. A similar approach is presented in the work of Kennedy and Inkpen [10], this time using an opinion lexicon based on the combination of other existing resources.
Manually built lexicons however tend to be constrained to a small number of terms. By its nature, building manual lists is a time consuming effort, and may be subject to annotator bias. To overcome these issues *lexical induction* approaches have been proposed in the literature with a view to extend the size of opinion lexicons from a core set of seed terms, either by exploring term relationships, or by evaluating similarities in document corpora. Early work in this area seen in [9] extends a list of positive and negative adjectives by evaluating conjunctive statements in a document corpus. Another common approach is to derive opinion terms from the WordNet database of terms and relationships [12], typically by examining the semantic relationships of a term such as synonyms and antonyms.

Lexicons built using this approach can be seen applied to subjectivity detection research in [21] and applied to sentiment classification in [4] and [16].

2.1 WordNet Glosses and SentiWordNet

As noted in [15], term relationships in the WordNet database form a highly disconnected graph, and thus expansion of opinion information from a core of seed words by examining semantic relationships such as synonyms and antonyms is bound to be restricted only to a subset of terms. To overcome this problem, information contained in term *glosses* – explanatory text accompanying each term – can be explored to infer term orientation, based on the assumption that a given term and the terms contained in its gloss are likely to indicate the same polarity. In [2] a method for lexicon expansion is proposed where terms are assigned positive or negative opinions based on the existence of terms known to carry opinion content found on the term gloss. The authors argue that glosses have a potentially low level of noise since they "are designed to match as close as possible the components of meaning of the word, have relatively standard style, grammar and syntactic structure"; This idea is also seen in [5], this time by using supervised learning methods for extending a lexicon by exploring gloss information, yielding positive accuracy improvements over a gold standard in comparison to some of the methods previously discussed in this section. This is the same approach employed on building the *SentiWordNet* opinion lexicon [6].
SentiWordNet is built in a two-stage approach: initially, WordNet term relationships such as synonym, antonym and hyponymy are explored to extend a core of seed words used in [19], and known a priori to carry positive or negative opinion bias. After a fixed number of iterations, a subset of WordNet terms is obtained with either a positive or negative label. These term's glosses are then used to train a committee of machine learning classifiers. To minimize bias, the classifiers are trained using different algorithms and different training set sizes. The predictions from the classifier committee are then used to determine the sentiment orientation of the remainder of terms in WordNet. The table below compares the coverage of SentiWordNet in relation to other manually built opinion lexicons available in the literature.

Opinion Lexicon	**Total Sentiment Bearing Terms**
General Inquirer [1] [17].	4216
Subjectivity Clues Lexicon [21].	7650 (out of 8221 terms)
Grefenstette et al [8].	2258
SentiWordNet [6].	28431 (out of total 86994 WordNet terms)

Table 1. Coverage of Opinion Lexicons

3 Approach

Our research assesses the use of SentiWordNet to the task of document level sentiment classification using the *Polarity* data set of film reviews presented in [14]. Initially, the lexicon was applied by counting positive and negative terms found in a document and determining sentiment orientation based on which class received the highest score, similar to the methods presented in [13] and [10]. A refinement to this method consisted on building a data set of features derived from SentiWordNet scores, following a careful evaluation of the data set and SentiWordNet.
Each set of terms sharing the same meaning in SentiWordNet (*synsets*) is associated with two numerical scores ranging from 0 to 1, each indicating the synset's positive and negative bias. The scores reflect the agreement amongst the classifier committee on the positive or negative label for a term, thus one distinct aspect of SentiWordNet is that it is possible for a term to have non-zero values for both positive and negative scores, according to the formula:

[1] *http://www.wjh.harvard.edu/~inquirer*

$$Pos.\ Score(term) + Neg.\ Score(term) + Objective\ Score(term) = 1 \quad (1)$$

Terms in the SentiWordNet database follow the categorization into parts of speech derived from WordNet, and therefore to correctly apply scores to terms, a part of speech tagger program was applied to the polarity data set. In our experiment, the *Stanford Part of Speech Tagger* described in [18] was used.

SentiWordNet scores were then calculated for terms found, and additional metrics were calculated from the scores. Overall scores for each part of speech were computed, along with ratios of scores in relation to number of terms. Documents were also divided into equally sized segments, and scoring was performed on each segment to assess the impact of different parts of the document to overall sentiment. A total of 96 distinct features were generated as summarized on the table below.

Metric Category	**Features**
Overall Document Scores	Sum of positive and negative scores for Adjectives. Sum of positive and negative scores for Adverbs. Sum of positive and negative scores for Verbs.
Score ratio to total terms	Ratio of overall score per total terms found, for each part of speech.
Positive to negative score ratios	Positive to negative scores ratio per part of speech.
Scores per document segment	Ratios for the above metrics for each of N partitions of a document. • Each document was segmented into 10 partitions with equal number of terms.
Negation	Percentage of negated terms in document.

Table 2. Metrics Derived From SentiWordNet

3.1 Natural Language and Style Considerations

Another aspect evaluated by this experiment was the influence of applying weights to scores as a function of its position in the document. This would intuitively translate to the existence of areas within a document that tend to carry more opinion content, such as the end of the document where closing remarks would reflect the general author view. Several adjusting schemes were attempted and the chosen method implements a linearly increasing weight adjustment to scores, as given by the formula below.

$$score_{adj} = score\,\frac{t_i}{T}C \quad (2)$$

With C being a constant value, and t_i the position of the given term t relative to the total of terms T in the document.

Negation detection is also an important element of implementing sentiment analysis by using term scores, since negation in a sentence such as “I did *not* find this movie funny or interesting” would invert the opinion orientation of otherwise positive terms such as “funny” and “interesting”. This research implemented a version of the *NegEx* algorithm [3] for negation detection, which scans sentences based on a database of pre defined negation expressions. The algorithm maintains three distinct lists, depending on the scope of the negation: expressions that modify preceding terms, subsequent terms and pseudo-negation expressions with no effect on term polarity.

Finally, the data set was generated from the source documents by extracting the above information with SentiWordNet. A support vector machine classifier was then trained based on a label indicating positive and negative sentiment, and classification performance was measured using average

accuracies and 3-fold cross validation. The experiment was executed using the support vector machine implementation available in the *RapidMiner* data mining application [11].

4 Results

4.1 Term Counting

SentiWordNet scores were calculated as positive and negative terms were found on each document, and used to determine sentiment orientation by assigning the document to the class with the highest score. This method yielded an overall accuracy of **65.85%,** with results detailed in the table below.

Class	**Positive**	**Negative**
Predicted Positive	576	259
Predicted Negative	424	741
Total	1000	1000
Class Recall	57.6%	74.1%
Class Precision	68.98%	63.76%

Table 3. SentiWordNet Score Counting Results

4.1 SentiWordNet Features

For this method, a linear support vector machine classifier was trained using the features derived from SentiWordNet detailed on Section 3. Best results were obtained when combined with a feature selection refinement step based on attribute information gain. The table below presents accuracies for each stage of the experiment. It can be noticed that small improvements were obtained when negation detection and scoring functions were added to the model.

Experiment	**Accuracy**
SentiWordNet Features (no refinement).	67.40%
- Including Linear Weight Adjustment to Scores.	68.00%
- Including Negation Detection and Linear Weight Scoring.	68.50%
SentiWordNet, Negation Detection, Linear Scoring and Feature Selection.	69.35%

Table 4. SVM Accuracy Results

5 Discussion

The table below illustrates how SentiWordNet compares to other published results in the area using the same data set and similar approaches bases on opinion lexicons.

Method	**Accuracy**
SentiWordNet – Term Counting (this research)	**65.85%**
SentiWordNet Scores used as Features (this research).	**69.35%**
Term Counting - Manually built list of Positive/Negative words [13].	69.00%
Term counting from Combined Lexicon and valence shifters [10].	67.80%

Table 5. Accuracy Comparisons

Term counting using SentiWordNet remains close to other results using manually built lexicons, which is encouraging for the use of resources built from semi supervised methods. Our second method using SentiWordNet as a source of features for a supervised learning algorithm yielded improvements

over the term counting approach. The use of weight adjustment has yielded small improvements to the method, suggesting remarks affecting overall sentiment being placed towards the end of a document.
On both cases, the results are within close range of other results employing opinion lexicons seen in the literature: In [13] the results are based on term counting from a manually built word list for the domain of film reviews, whereas results from [10] follow the same principle, but leverage a combined lexicon and take into account *intensifier* and *diminisher* terms such as “very” and “seldom”.

4.1 Misclassifications

Results for the term counting approach seen in Table 3 show that the method provides better recall for the negative class than the positive one. This may indicate a stronger and more explicit choice of terms on negative reviews than in positive ones, and that authors are more likely to include negative remarks on positive reviews for a more balanced assessment, like the ones seen in the concluding remarks of a film review presented below:

“the only downfall of the opening sequence is the editing style used… it’s choppy, slow motion which is unsettling and distracting.”

The phenomenon of *thwarted expectations* reported in [13] can also affect this method, where the author chooses to build up the expectation of a good film, for example by mentioning director and actor’s previous achievements, only to later frustrate it by presenting an overall negative view. On those cases, the number of terms with positive orientation would be high, therefore affecting conclusions made by a classifier using data based on term polarity.
Some inaccuracies seen on SentiWordNet scores may be caused by the reliance on glosses as a source of information for determining term orientation. As an example the term *ludicrous* has a positive score in SentiWordNet, and the following gloss:

“absurd, cockeyed, derisory, idiotic, laughable, ludicrous, nonsensical, preposterous, ridiculous (incongruous; inviting ridicule) "the absurd excuse that the dog ate his homework"; "that's a cockeyed idea"; "ask a nonsensical question and get a nonsensical answer"; "a contribution so small as to be laughable"; "it is ludicrous to call a cottage a mansion"; "a preposterous attempt to turn back the pages of history"; "her conceited assumption of universal interest in her rather dull children was ridiculous."

It can be argued that this term should contain a negative orientation, given its association to the synonyms *farcical* and *idiotic*. However SentiWordNet may have chosen a positive score on the basis the gloss text is more likely to be associated with a positive term than a negative one: terms such as *exuberance* and *clown* and the somewhat ambiguous *laughable* could be influencing the construction method in assigning incorrect scores. The dependence of SentiWordNet scores on term glosses could be a limiting factor in the accuracy of term scores and the overall classification accuracy of this method.
Finally, the use of colloquial language and expressions where no opinion information exists, disambiguation of WordNet terms with more than one meaning, inaccuracies in the assignment of part of speech tags, and the correct detection of named entities such as actor and film names were identified as contributing factors to misclassifications seen using this method.

5 Conclusions and Future Work

This research assessed the use of the SentiWordNet opinion lexicon in the task of sentiment classification of film reviews. Results obtained by simple word counting were similar to other results employing manual lexicons, indicating SentiWordNet performs well when compared with manual resources on this task. In addition, using SentiWordNet as a source of features for a supervised learning scheme has shown improvements over pure term counting. This study also revealed opportunities where further linguistic processing yield gains in classification accuracies. These,

coupled with the relative low dimensionality of a data set built from SentiWordNet data set - less than 100 features compared to several thousand typically seen on word vector approaches - could lead to more attractive models for real world applications.
Further aspects of our research will involve a more detailed comparison of the performance of SentiWordNet and other lexicons on similar opinion mining tasks could help in better understanding their strengths, and how they can be used together. This could be particularly beneficial in overcoming some of the limitations seen in SentiWordNet's reliance on glosses. In addition, research in combining a classifier based on SentiWordNet with other approaches such as word vectors may produce better results than each individual classifier can produce on its own. Some encouraging empirical results of such methods applied to sentiment classification research are seen in [9] and [23].

Acknowledgements

We wish to thank Andrea Esuli and Fabrizio Sebastiani, from the Italian Institute of Information Science and Technology, for making the SentiWordNet lexical resource available for use in this research.

References

[1] 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*, 26, 3 (Jun. 2008), 1-34.

[2] Andreevskaya A., Bergler S. (2006). Mining WordNet for Fuzzy Sentiment: Sentiment Tag Extraction from WordNet Glosses. *In Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics – EACL 2006*.

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

[4] Dave K, Lawrence S, Pennock D. (2003). 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*, (May 20-24, 2003), Budapest, Hungary.

[5] Esuli, A. and Sebastiani, F. (2005). Determining the semantic orientation of terms through gloss classification. *Proceedings of the 14th ACM international Conference on information and Knowledge Management* (Bremen, Germany, October 31 - November 05, 2005). CIKM '05. ACM, New York, NY, 617-624.

[6] Esuli A, Sebastiani F. (2006). SentiWordNet: A Publicly Available Lexical Resource for Opinion Mining. *Proceedings from International Conference on Language Resources and Evaluation (LREC)*, Genoa, 2006.

[7] Gamon, M. (2004). Sentiment Classification on Customer Feedback Data: Noisy Data, Large Feature Vectors, and the Role of Linguistic Analysis. *Proceedings of the 20th international conference on Computational Linguistics*. Geneva, Switzerland: Association for Computational Linguistics.

[8] Grefenstette G., Qu Y., Shanahan J., Evans d. (2004). Coupling Niche Browsers and Affect Analysis for an Opinion Mining Application. *Proceedings of the RIAO 2004*, pp.186-194.

[9] Hatzivassiloglou, V., and McKeown, K. (1997). Predicting the Semantic Orientation of Adjectives. *Preceedings of the 35th Annual Meeting of the Association of Computational Linguistics (ACL'97)*. Madrid, Spain, pp. 174-181.

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

[11] Mierswa I., Wurst M., Klinkenberg R., Scholz M., Euler T. (2006). YALE: Rapid Prototyping for Complex Data Mining Tasks. *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-06)*.

[12] Miller G. A., Beckwith R., Fellbaum C, Gross D, Miller K. J. (1990). Introduction to Wordnet: An On-line Lexical Database. *International Journal of Lexicography.* Vol. 3, No. 4 (Jan. 1990), 235-244.

[13] Pang B., Lee L., and Vaithyanathan, S. (2002). Thumbs up? Sentiment Classification using Machine Learning Techniques. *Proceedings of EMNLP*, 2002.

[14] Pang B., Lee L. (2004). A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts. *Proceedings of the ACL*, 2004.

[15] Rao D. and Ravichandran D. (2009). Semi-Supervised Polarity Lexicon Induction. *Proceedings of the 12th Conference of the European Chapter of the ACL*. Athens, Greece (2009, Mar. 30th to Apr. 3rd), 675-682.

[16] Salvetti F., Lewis S., Reichenbach C. (2004). Automatic Opinion Polarity Classification of Movie Reviews. *Colorado Research in Linguistics*. Volume 17, Issue 1 (June 2004). Boulder: University of Colorado.

[17] Stone, P.J., Dunphy, D.C., Smith,M.S., Oglivie D.M. (1966). The General Enquirer: A computer Approach to Content Analysis. *MIT Press*, Cambridge MA.

[18] Toutanova K., Manning C. (2000). Enriching the Knowledge Sources Used in a Maximum Entropy Part-of-Speech Tagger. *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-2000)*, pp. 63-70.

[19] Turney P., and Littman M. (2003). Measuring praise and Criticism: Inference of Semantic Orientation from Association. *ACM Transactions on Information Systems*, No. 21, 4, 315–346.

[20] Turney P. (2002). Thumbs up or Thumbs down? Sentiment Orientation Applied to Unsupervised Classification of Reviews. *Proceedings of the 40th Annual Meeting of the Association of Computational Linguistics – ACL*, 2002.

[21] Wilson T., Wiebe J., and Hoffmann P. (2005). Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis. *Proceedings of HLT/EMNLP*, Vancouver, Canada.

[22] Yang K., Yu N., Zhang H. (2007). WIDIT in TREC-2007 Blog Track: Combining Lexicon-based Methods to Detect Opinionated Blogs. *Proceedings of the 16th Text Retrieval Conference (TREC 2007).*

[23] Yu H., Hatzivassiloglou V. (2003). Towards Answering Opinion Questions: Separating Facts from Opinions and Identifying Polarity in Sentences. *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, 129-136.