Sentiment analysis using support vector machines with diverse information sources

Tony Mullen and Nigel Collier
National Institute of Informatics (NII)
Hitotsubashi 2-1-2, Chiyoda-ku
Tokyo 101-8430
Japan
{mullen,collier}@nii.ac.jp

Abstract

This paper introduces an approach to sentiment analysis which uses support vector machines (SVMs) to bring together diverse sources of potentially pertinent information, including several favorability measures for phrases and adjectives and, where available, knowledge of the topic of the text. Models using the features introduced are further combined with unigram models which have been shown to be effective in the past (Pang et al., 2002) and lemmatized versions of the unigram models. Experiments on movie review data from Epinions.com demonstrate that hybrid SVMs which combine unigram-style feature-based SVMs with those based on real-valued favorability measures obtain superior performance, producing the best results yet published using this data. Further experiments using a feature set enriched with topic information on a smaller dataset of music reviews hand-annotated for topic are also reported, the results of which suggest that incorporating topic information into such models may also yield improvement.

1 Introduction

Recently an increasing amount of research has been devoted to investigating methods of recognizing favorable and unfavorable sentiments towards specific subjects within natural language texts. Areas of application for such analysis are numerous and varied, ranging from newsgroup flame filtering and informative augmentation of search engine responses to analysis of public opinion trends and customer feedback. For many of these tasks, classifying the tone of the communication as generally positive or negative is an important step.

There are a number of challenging aspects of this task. Opinions in natural language are very often expressed in subtle and complex ways, presenting challenges which may not be easily addressed by simple text categorization approaches such as n-gram or keyword identification approaches. Although such approaches have been employed effectively (Pang et al., 2002), there appears to remain considerable room for improvement. Moving beyond these approaches can involve addressing the task at several levels. Recognizing the semantic impact of words or phrases is a challenging task in itself, but in many cases the overarching sentiment of a text is not the same as that of decontextualized snippets. Negative reviews may contain many apparently positive phrases even while maintaining a strongly negative tone, and the opposite is also common.

This paper introduces an approach to classifying texts as positive or negative using Support Vector Machines (SVMs), a well-known and powerful tool for classification of vectors of real-valued features (Vapnik, 1998). The present approach emphasizes the use of a variety of diverse information sources, and SVMs provide the ideal tool to bring these sources together. We describe the methods used to assign values to selected words and phrases, and we introduce a method of bringing them together to create a model for the classification of texts. In addition, several classes of features based upon the proximity of the topic with phrases which have been assigned favorability values are described in order to take further advantage of situations in which the topic of the text may be explicitly identified. The results of a variety of experiments are presented, using both data which is not topic annotated and data which has been hand annotated for topic. In the case of the former, the present approach is shown to yield better performance than previous models on the same data. In the case of the latter, results indicate that our approach may allow for further improvements to be gained given knowledge of the topic of the text.

2 Motivation

A continual challenge in the task of sentiment analysis of a text is to home in on those aspects of the text which are in some way representative of the tone of the whole text. In the past, work has been done in the area of characterizing words and

phrases according to their emotive tone (Turney and Littman, 2003; Turney, 2002; Kamps et al., 2002; Hatzivassiloglou and Wiebe, 2000; Hatzivassiloglou and McKeown, 2002; Wiebe, 2000), but in many domains of text, the values of individual phrases may bear little relation to the overall sentiment expressed by the text. Pang et al. (2002)'s treatment of the task as analogous to topic-classification underscores the difference between the two tasks. Sources of misleading phrases include what Pang et al. (2002) refer to as "thwarted expectations" narrative, where emotive effect is attained by emphasizing the contrast between what the reviewer expected and the actual experience. For example, in the record review data used in the present experiments, the sentence, "How could they not be the most unimaginative, bleak, whiny emo band since..." occurs in one of the most highly rated reviews, describing the reviewer's initial misgivings about the record under review based on its packaging, followed immediately by "I don't know. But it's nothing like you'd imagine. Not even almost." Clearly, the strongly positive sentiment conveyed by these four sentences is much different from what we would expect from the sum of its parts. Likewise, another exceptionally highly rated review contains the quote: "This was a completely different band, defeated, miserable, and exhausted, absolutely, but not hopeless: they had somehow managed to succeed where every other band in their shoes had failed." Other rhetorical devices which tend to widen the gap in emotional tone between what is said locally in phrases and what is meant globally in the text include the drawing of contrasts between the reviewed entity and other entities, sarcasm, understatement, and digressions, all of which are used in abundance in many discourse domains.

The motivation of the present research has been to incorporate methods of measuring the favorability content of phrases into a general classification tool for texts.

3 Methods

3.1 Semantic orientation with PMI

Here, the term *semantic orientation* (SO) (Hatzivassiloglou and McKeown, 2002) refers to a real number measure of the positive or negative sentiment expressed by a word or phrase. In the present work, the approach taken by Turney (2002) is used to derive such values for selected phrases in the text. This approach is simple and surprisingly effective. Moreover, is not restricted to words of a particular part of speech, nor even restricted to single words, but can be used with multiple word phrases. In general, two word phrases conforming to particular part-of-speech templates representing possible descriptive combinations are used. The phrase patterns used by Turney can be seen in figure 1. In some cases, the present approach deviates from this, utilizing values derived from single words. For the purposes of this paper, these phrases will be referred to as *value phrases*, since they will be the sources of SO values. Once the desired value phrases have been extracted from the text, each one is assigned an SO value. The SO of a phrase is determined based upon the phrase's *pointwise mutual information* (PMI) with the words "excellent" and "poor". PMI is defined by Church and Hanks (1989) as follows:

$$\mathrm{PMI}(w_1, w_2) = \log_2 \left(\frac{p(w_1 \ \& \ w_2)}{p(w_1)p(w_2)} \right) \qquad (1)$$

where $p(w_1 \ \& \ w_2)$ is the probability that w_1 and w_2 co-occur.

The SO for a $phrase$ is the difference between its PMI with the word "excellent" and its PMI with the word "poor." The probabilities are estimated by querying the AltaVista Advanced Search engine[1] for counts. The search engine's "NEAR" operator, representing occurrences of the two queried words within ten words of each other in a text, is used to define co-occurrence. The final SO equation is

$$\mathrm{SO}(phrase) = log_2 \left(\frac{\mathrm{hits}(phrase \ \mathrm{NEAR} \text{“excellent”}) \mathrm{hits}(\text{“poor”})}{\mathrm{hits}(phrase \ \mathrm{NEAR} \text{“poor”}) \mathrm{hits}(\text{“excellent”})} \right)$$

Intuitively, this yields values above zero for phrases with greater PMI with the word "excellent" and below zero for greater PMI with "poor". A SO value of zero would indicate a completely neutral semantic orientation.

3.2 Osgood semantic differentiation with WordNet

Further feature types are derived using the method of Kamps and Marx (2002) of using WordNet relationships to derive three values pertinent to the emotive meaning of adjectives. The three values correspond to the *potency* (strong or weak), *activity* (active or passive) and the *evaluative* (good or bad) factors introduced in Charles Osgood's Theory of Semantic Differentiation (Osgood et al., 1957).

[1] www.altavista.com

	First Word	Second Word	Third Word (Not Extracted)
1.	JJ	NN or NNS	anything
2.	RB, RBR, or RBS	JJ	not NN nor NNS
3.	JJ	JJ	not NN nor NNS
4.	NN or NNS	JJ	not NN or NNS
5.	RB, RBR, or RBS	VB, VBD, VBN or VBG	anything

Figure 1: Patterns for extraction of value phrases in Turney (2002)

These values are derived by measuring the relative minimal path length (MPL) in WordNet between the adjective in question and the pair of words appropriate for the given factor. In the case of the *evaluative* factor (EVA) for example, the comparison is between the MPL between the adjective and "good" and the MPL between the adjective and "bad".

Only adjectives connected by synonymy to each of the opposites are considered. The method results in a list of 5410 adjectives, each of which is given a value for each of the three factors referred to as EVA, POT, and ACT. For the purposes of this research, each of these factors' values are averaged over all the adjectives in a text, yielding three real-valued feature values for the text, which will be added to the SVM model.

3.3 Topic proximity and syntactic-relation features

Our approach shares the intuition of Natsukawa and Yi (2003) that sentiment expressed with regard to a particular subject can best be identified with reference to the subject itself. Collecting emotive content from a text overall can only give the most general indication of the sentiment of that text towards the specific subject. Nevertheless, in the present work, it is assumed that the pertinent analysis will occur at the text level. The key is to find a way to incorporate pertinent semantic orientation values derived from phrases into a model of texts. Our approach seeks to employ semantic orientation values from a variety of different sources and use them to create a feature space which can be separated into classes using an SVM.

In some application domains, it is known in advance what the topic is toward which sentiment is to be evaluated. The present approach allows for the incorporation of features which exploit this knowledge, where available. This is done by creating several classes of features based upon the semantic orientation values of phrases given their position in relation to the topic of the text.

Although in opinion-based texts there is generally a single primary subject about which the opinion is favorable or unfavorable, it would seem that secondary subjects may also be useful to identify. The primary subject of a book review, for example, is a book. However, the review's overall attitude to the author may also be enlightening, although it is not necessarily identical to the attitude towards the book. Likewise in a product review, the attitude towards the company which manufactures the product may be pertinent. It is an open question whether such secondary topic information would be beneficial or harmful to the modeling task. The approach described in this paper allows such secondary information to be incorporated, where available.

In the second of the two datasets used in the present experiments, texts were annotated by hand using the Open Ontology Forge annotation tool (Collier et al., 2003). In each record review, references (including co-reference) to the record being reviewed were tagged as `THIS_WORK` and references to the artist under review were tagged as `THIS_ARTIST`.

With these entities tagged, a number of classes of features may be extracted, representing various relationships between topic entities and value phrases similar to those described in section 3.1. The classes looked at in this work are as follows:

Turney Value The average value of all value phrases' SO values for the text. Classification by this feature alone is not the equivalent of Turney's approach, since the present approach involves retraining in a supervised model.

In sentence with THIS_WORK The average value of all value phrases which occur in the same sentence as a reference to the work being reviewed.

Following THIS_WORK The average value of all value phrases which follow a reference to the work being reviewed directly, or separated only by the copula or a preposition.

Preceding THIS_WORK The average value of all value phrases which precede a reference to

the work being reviewed directly, or separated only by the copula or a preposition.

In sentence with THIS_ARTIST As above, but with reference to the artist.

Following THIS_ARTIST As above, but with reference to the artist.

Preceding THIS_ARTIST As above, but with reference to the artist.

The features used which make use of adjectives with WordNet derived Osgood values include the following:

Text-wide EVA The average EVA value of all adjectives in a text.

Text-wide POT The average POT value of all adjectives in a text.

Text-wide ACT The average ACT value of all adjectives in a text.

TOPIC-sentence EVA The average EVA value of all adjectives which share a sentence with the topic of the text.

TOPIC-sentence POT The average POT value of all adjectives which share a sentence with the topic of the text.

TOPIC-sentence ACT The average ACT value of all adjectives which share a sentence with the topic of the text.

The grouping of these classes should reflect some common degree of reliability of features within a given class, but due to data sparseness what might have been more natural class groupings—for example including **value-phrase_preposition_topic-entity** as a distinct class—often had to be conflated in order to get features with enough occurrences to be representative.

For each of these classes a value may be derived for a text. Representing each text as a vector of these real-valued features forms the basis for the SVM model. In the case of data for which no explicit topic information is available, only the Turney value is used from the first list, and the Text-wide EVA, POT, and ACT values from the second list. A resultant feature vector representing a text may be composed of a combination of boolean unigram-style features and real-valued favorability measures in the form of the Osgood values and the PMI derived values.

3.4 Support Vector Machines

SVMs are a machine learning classification technique which use a function called a *kernel* to map a space of data points in which the data is not linearly separable onto a new space in which it is, with allowances for erroneous classification. For a tutorial on SVMs and details of their formulation we refer the reader to Burges (1998) and Cristiani and Shawe-Tailor (2000). A detailed treatment of these models' application to text classification may be found in Joachims (2001).

4 Experiments

First, value phrases were extracted and their values were derived using the method described in section 3.1. After this, supervised learning was performed using these values as features. In training data, reviews corresponding to a below average rating were classed as negative and those with an above average rating were classed as positive.

The first dataset consisted of a total of 1380 Epinions.com movie reviews, approximately half positive and half negative. This is the same dataset as was presented in Pang et al.(2002). In order to compare results as directly as possible, we report results of 3-fold cross validation, following Pang et al.(2002). Likewise, we include punctuation as tokens and normalize the feature values for text length. To lend further support to the conclusions we also report results for 10-fold cross validation experiments. On this dataset the feature sets investigated include various combinations of the Turney value, the three text-wide Osgood values, and word token unigrams or lemmatized unigrams. [2]

The second dataset consists of 100 record reviews from the Pitchfork Media online record review publication,[3] topic-annotated by hand. In addition to the features employed with the first dataset, this dataset allows the use those features described in 3.3 which make use of topic information, namely the broader PMI derived SO values and the topic-sentence Osgood values. Due to the relatively small size of this dataset, test suites were created using 100, 20, 10, and 5-fold cross validation, to maximize the amount of data available for training and the accuracy of the results. Text length normalization appeared to harm performance on this dataset, and so the models reported here for this dataset were not normalized for length.

SVMs were built using Kudo's TinySVM soft-

[2] We employ the Conexor FDG parser (Tapanainen and Järvinen, 1997) for POS tagging and lemmatization

[3] `http://www.pitchforkmedia.com`

Model	3 folds	10 folds
Pang et al. 2002	82.9%	NA
Turney Values only	68.4%	68.3%
Osgood only	56.2%	56.4%
Turney Values and Osgood	69.0%	68.7%
Unigrams	82.8%	83.5%
Unigrams and Osgood	82.8%	83.5%
Unigrams and Turney	83.2%	85.1%
Unigrams, Turney, Osgood	82.8%	85.1%
Lemmas	84.1%	85.7%
Lemmas and Osgood	83.1 %	84.7%
Lemmas and Turney	84.2%	84.9%
Lemmas, Turney, Osgood	83.8%	84.5%
Hybrid SVM (Turney and Lemmas)	84.4%	**86.0%**
Hybrid SVM (Turney/Osgood and Lemmas)	**84.6%**	**86.0%**

Figure 2: Accuracy results for 3 and 10-fold cross-validation tests on Epinions.com movie review data using a linear kernel.

ware implementation.[4] Several kernel types, kernel parameters, and optimization parameters were investigated, but no appreciable and consistent benefits were gained by deviating from the the default linear kernel with all parameter values set to their default, so only these results are reported here, with the exception of the Turney Values-only model on the Pitchfork dataset. This single-featured model caused segmentation faults on some partitions with the linear kernel, and so the results for this model only, seen in figure 3, were obtained using a polynomial kernel with parameter d set to 2 (default is 1) and the constraints violation penalty set at 2 (default is 1).

Several *hybrid* SVM models were further tested using the results from the previously described models as features. In these models, the feature values for each event represent the distance from the dividing hyperplane for each constituent model.

5 Results

The accuracy value represents the percentage of test texts which were classified correctly by the model. Results on the first dataset, without topic information, are shown in figure 2. The results for 3-fold cross validation show how the present feature sets compare with the best performing SVM reported in Pang et al.

In general, the addition of Osgood values does not seem to yield improvement in any of the models. The Turney values appear more helpful, which is not surprising given their superior performance alone. In the case of the SVM with only a single Turney value, accuracy is already at 68.3% (Turney (2002) reports that simply averaging these values on the same data yields 65.8% accuracy). The Osgood values are considerably less reliable, yielding only 56.2% accuracy on their own. Lemmas outperform unigrams in all experiments, and in fact the simple lemma models outperform even those augmented with the Turney and Osgood features in the experiments on the epinions data. The contribution of these new feature types is most pronounced when they are used to train a separate SVM and the two SVMs are combined in a hybrid SVM. The best results are obtained using such hybrid SVMs, which yield scores of 84.6% accuracy on the 3-fold experiments and 86.0% accuracy on the 10-fold experiments.

In the second set of experiments, again, inclusion of Osgood features shows no evidence of yielding any improvement in modeling when other features are present, indeed, as in the previous experiments there are some cases in which these features may be harming performance. The PMI values, on the other hand, appear to yield consistent improvement. Furthermore on both the 20 and 100-fold test suites the inclusion of all PMI values with lemmas outperforms the use of only the Turney values, suggesting that the incorporation of the available topic relations is helpful. Although there is not enough data here to be certain of trends, it is intuitive that the broader PMI values, similarly to the unigrams, would particularly benefit from increased training data, due to their specificity, and therefore their relative sparse-

[4] `http://cl.aist-nara.ac.jp/~taku-ku/software/TinySVM`

Model	5 folds	10 folds	20 folds	100 folds
Turney Values only	72%	73%	72%	72%
All (THIS_WORK and THIS_ARTIST) PMI	70%	70%	68%	69%
THIS_WORK PMI	72%	69%	70%	71%
All Osgood	64%	64%	65%	64%
All PMI and Osgood	74%	71%	74%	72%
Unigrams	79%	80%	78%	82%
Unigrams, PMI, Osgood	81%	80%	82%	82%
Lemmas	83%	85%	84%	84%
Lemmas and Osgood	83%	84%	84%	84%
Lemmas and Turney	84%	85%	84%	84%
Lemmas, Turney, text-wide Osgood	84%	85%	84%	84%
Lemmas, PMI, Osgood	84%	85%	84%	86%
Lemmas and PMI	84%	85%	**85%**	86%
Hybrid SVM (PMI/Osgood and Lemmas)	**86%**	**87%**	84%	**89%**

Figure 3: Accuracy results for 5, 10, 20 and 100-fold cross-validation tests with Pitchforkmedia.com record review data, hand-annotated for topic. Note that the results for the Turney Values-only model were obtained using a polynomial kernel. All others were obtained with a linear kernel.

ness. Once again, the information appears to be most fruitfully combined by building SVMs representing semantic values and lemmas separately and combining them in a single hybrid SVM. The average score over the four n-fold cross validation experiments for the hybrid SVM is 86.5%, whereas the average score for the second-best performing model, incoporating all semantic value features and lemmas, is 85%. The simple lemmas model obtains an average score of 84% and the simple unigrams model obtains 79.75%.

6 Discussion

The main development presented here is the incorporation of several new information sources as features into SVMs which previously relied entirely on the effective but limited "bag of words" approach. The ability of SVMs to handle real-valued features makes this possible, and the information sources introduced in the work Turney and Kamps and Marx provide sensible places to start. The intuition that topic relations and proximity should also yield gains also appears to be borne out in the present experiments. The various sources of information appear to be best combined by integrating several distinct SVMs.

6.1 Other issues

At the level of the phrasal SO assignment, it would seem that some improvement could be gained by adding domain context to the AltaVista Search. Many—perhaps most—terms' favorability content depends to some extent on their context. As Turney notes, "unpredictable," is generally positive when describing a movie plot, and negative when describing an automobile or a politician. Likewise, such terms as "devastating" might be generally negative, but in the context of music or art may imply an emotional engagement which is usually seen as positive. Likewise, although "excellent" and "poor" as the poles in assessing this value seems somewhat arbitrary, cursory experiments in adjusting the search have thus far supported Turney's conclusion that the former are the appropriate terms to use for this task. One problem with limiting the domain by adding topic-related word constraints to the query is that the resultant hit count is greatly diminished, canceling out any potential gain. It is to be hoped that in the future, as search engines continue to improve and the Internet continues to grow, more possibilities will open up in this regard.

It also seems likely that the topic-relations aspect of the present research only scratches the surface of what should be possible. There is still considerable room for improvement in performance. The present models may also be further expanded with features representing other information sources, which may include other types of semantic annotation (Wiebe, 2002), or features based on more sophisticated grammatical or dependency relations or on zone information. In any case, it is hoped that the present work may help to indicate how various information sources pertinent to the task may be brought together.

7 Conclusion

The method introduced in this paper allows several methods of assigning semantic values to phrases and words within a text to be exploited in a more useful way than was previously possible, by incorporating them as features for SVM modeling, and for explicit topic information to be utilized, when available, by features incorporating such values. Combinations of SVMs using these features in conjunction with SVMs based on unigrams and lemmatized unigrams are shown to outperform models which do not use these information sources. The approach presented here is flexible and suggests promising avenues of further investigation.

References

C. Burges. 1998. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2(2):121–167.

K.W. Church and P. Hanks. 1989. Word association norms, mutual information and lexicography. In *Proceedings of the 27th Annual Conference of the ACL*, New Brunswick, NJ.

N. Collier, K. Takeuchi, A. Kawazoe, T. Mullen, and T. Wattarujeekrit. 2003. A framework for integrating deep and shallow semantic structures in text mining. In *Proceedings of the Seventh International Conference on Knowledge-based Intelligent Information and Engineering Systems*. Springer-Verlag.

N. Cristianini and J. Shawe-Taylor. 2000. *An Introduction to Support Vector Machines and other Kernel-based Learning Methods*. Cambridge University Press.

V. Hatzivassiloglou and K.R. McKeown. 2002. Predicting the semantic orientation of adjectives. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and the 8th Conference of the European Chapter of the ACL*.

V. Hatzivassiloglou and J. Wiebe. 2000. Effects of adjective orientation and gradability on sentence subjectivity.

Thorsten Joachims. 2001. *Learning to Classify Text Using Support Vector Machines*. Kluwer Academic Publishers.

Jaap Kamps, Maarten Marx, Robert J. Mokken, and Marten de Rijke. 2002. Words with attitude. In *In Proceedings of the 1st International Conference on Global WordNet*, Mysore, India.

Tetsuya Nasukawa and Jeonghee Yi. 2003. Sentiment analysis: Capturing favorability using natural language processing. In *Second International Conference on Knowledge Capture*, Florida, USA.

Charles E. Osgood, George J. Succi, and Percy H. Tannenbaum. 1957. *The Measurement of Meaning*. University of Illinois.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? Sentiment classification using machine learning techniques. In *Empirical Methods in Natural Language Processing [and Very Large Corpora]*.

P. Tapanainen and T. Järvinen. 1997. A non-projective dependency parser. In *Proceedings of the 5th Conference on Applied Natural Language Processing, Washington D.C., Association of Computational Linguistics*.

P.D. Turney and M.L. Littman. 2003. Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems (TOIS)*, 21(4):315–346.

P.D. Turney. 2002. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, Philadelphia.

Vladimir Vapnik. 1998. *Statistical Learning Theory*. Wiley, Chichester, GB.

Janyce Wiebe. 2000. Learning subjective adjectives from corpora. In *Proc. 17th National Conference on Artificial Intelligence (AAAI-2000)*, Austin, Texas, July.

J Wiebe. 2002. Instructions for annotating opinions in newspaper articles. Technical Report TR-02-101, University of Pittsburgh, Pittsburgh, PA.