Weakly Supervised Techniques for Domain-Independent Sentiment Classification

Jonathon Read
Department of Informatics
University of Sussex
Falmer, Brighton, BN1 9QH
j.l.read@sussex.ac.uk

John Carroll
Department of Informatics
University of Sussex
Falmer, Brighton, BN1 9QH
j.a.carroll@sussex.ac.uk

ABSTRACT

An important sub-task of sentiment analysis is polarity classification, in which text is classified as being positive or negative. Supervised machine learning techniques can perform this task very effectively. However, they require a large corpus of training data, and a number of studies have demonstrated that the good performance of supervised models is dependent on a good match between the training and testing data with respect to the domain, topic and time-period.

Weakly-supervised techniques use a large collection of unlabelled text to determine sentiment, and so their performance may be less dependent on the domain, topic and time-period represented by the testing data. This paper presents experiments that investigate the effectiveness of word similarity techniques when performing weakly-supervised sentiment classification. It also considers the extent to which the performance of each method is independent from the domain, topic and time-period of the testing data. The results indicate that the word similarity techniques are suitable for applications that require sentiment classification across several domains.

Categories and Subject Descriptors

I.2.7 [**Artificial Intelligence**]: Natural Language Processing—*text analysis.*

General Terms

Algorithms, Experimentation.

Keywords

Distributional similarity, Lexical association, Semantic spaces, Sentiment analysis

1. INTRODUCTION

In recent years the World Wide Web has allowed both traditional publishers and the general public to distribute written content on a scale not previously possible. Newspapers reproduce much of their content online, while the blogosphere enables Web users to easily publish their thoughts for the consideration of the community. There are numerous professional and enthusiast review websites, and many online retailers, such as Amazon and iTunes, encourage their customers to review their purchases for the benefit of other shoppers. There is such a wealth of product reviews, in fact, that it has prompted the development of opinion-aggregation websites such as Metacritic.com.

This abundance of opinion is of interest to governments, companies and individuals seeking to distill public opinion for a variety of applications. For example, online retailing and opinion-aggregating websites could summarise a collection of reviews with an average sentiment score. This same task is of benefit to stock market traders who currently employ manual analyses of sentiment in financial news articles in order to predict stock fluctuations. Social network analyses are often driven by the frequencies of citations, but could be augmented by automatically determining whether such citations are positive or negative.

Other applications require more detail about the types of opinions expressed. For example, political parties and governmental departments are often interested in understanding public opinion on contentious issues, ans so commission person-to-person surveys. Similarly, traditional business market research techniques involve conducting surveys or organising focus group sessions to collect the opinions of a small number of members of the public. Instead, these tasks could be accomplished automatically by combining information retrieval techniques with opinion-mining to determine facets of opinion-bearing expressions such as its holder, target and nature [40].

Determining the polarity of a unit of text (whether it is generally positive or generally negative) is an important subtask when carrying out an analysis of sentiment. This is often achieved using supervised machine learning techniques which employ a large training corpus of labelled data [12, 17, 26, 31]. These models are very effective when classifying the sentiment of movie reviews, for example, typically attaining accuracies of around 86% (50% baseline) [30].

However, the performance of these supervised techniques is dependent on the degree to which the training and testing data match with respect to domain [3, 33], topic [10, 33] and time-period [33]. For instance, the accuracy of Support Vector Machine classifiers can fall by a mean of 6.5 percentage points when trained and tested on different topics in the domain of financial news, a mean of 16.5 percentage

TSA'09, November 6, 2009, Hong Kong, China.
Copyright 2009 ACM 978-1-60558-805-6/09/11 ...$10.00.

points when trained and tested on the different domains of newswire articles and movie reviews, and a mean of 4.2 percentage points when trained and tested movie reviews from different periods of time [33].

SO-PMI-IR [38] is an alternative method for sentiment classification that does not rely on labelled data. Its performance may therefore be vary less when testing on different data. This technique classifies a text by first extracting bigrams using certain part-of-speech patterns. A polarity score for each bigram is then calculated as its Pointwise Mutual Information (PMI, an information theoretic measure of lexical association) with a positive word *(excellent)* minus the PMI with a negative word *(poor)*, where PMI is estimated using document hit counts obtained from queries to the AltaVista search engine. The sentiment of a review is taken to be the sign of the sum of the sentiment scores for each extracted bigram; it is negative in sentiment if less than zero, or positive if greater than zero. SO-PMI-IR has been shown to be effective across four different domains of product reviews (automobiles, banks, movies and travel destinations), achieving an average accuracy of around 74% [38].

SO-PMI-IR is based on the notion that the polarity of a document can be derived from measurements of similarity between its constituent words and prototypical examples of sentiment. This paper presents experiments that investigate the effectiveness of pointwise mutual information and two other word similarity techniques when performing such measurements. It also considers the extent to which the performance of each method is independent from the domain, topic and time-period of the testing data.

Section 2 describes an adaptation of SO-PMI-IR that can employ any word similarity measure for sentiment classification and details three such measures: lexical association (using PMI), semantic spaces and distributional similarity. Section 3 presents experiments that investigate the effectiveness of these methods when constructing lexicons of polarity, scoring sentences according to the strength of sentiment and classifying movie reviews. Section 4 considers the degree to which the word similarity methods perform independent of domain, topic and time-period. Related work in the areas of domain-independent sentiment analysis and weakly-supervised sentiment analysis is presented in Section 5, while Section 6 offers conclusions and some directions for future work.

2. WORD SIMILARITY MEASURES FOR SENTIMENT ANALYSIS

Our method follows SO-PMI-IR in estimating the sentiment of a text as the sum of the sentiment of its constituents, except that all features in the text contribute to the score (rather than the extracted bigrams employed in the original study [37]). The method chooses the maximal scoring class c from a set of classes, $\mathbf{C}$, for a vector of features, $\mathbf{W}$:

$$\mathit{score}\left(\mathbf{W}, c\right) = \sum_{w \in \mathbf{W}} \frac{\sum_{p \in c_p} \mathrm{sim}\left(w, p\right)}{|c_p|} \tag{1}$$

where c_p is a set of prototypical example words of a class and $\mathrm{sim}\left(w, p\right)$ is some function that measures the semantic similarity of a word and a prototype. These methods are therefore *weakly*-supervised, as apposed to completely unsupervised, as they are provided with a basic definitions of the classes in the form of prototypical words.

We consider three methods of comparing word similarity: determining the strength of collocation through measures of lexical association; comparison of word's context using semantic spaces; and comparison of the dependency relations in which words appear using distributional similarity.

2.1 Lexical Association

Lexical association measures examine the first-order similarity between words [14]. That is, they determine the similarity of a pair of words by considering how likely they are to occur near each other. Pointwise Mutual Information [7] is one such measure. Following the notation above, it is defined as:

$$\mathrm{sim}_{pmi}\left(w, p\right) = \log_2 \frac{P\left(w, p\right)}{P\left(w\right) P\left(p\right)} \tag{2}$$

where the probabilities of seeing w and p jointly (that is, within some window of co-occurrence) or independently are typically estimated using frequencies observed in a corpus. PMI has been demonstrated to be an effective measure of semantic similarity. For instance, a method employing PMI achieved a score of 74% when answering synonym questions from the Test of English as a Foreign Language [37].

Pointwise Mutual Information is one of several measures of lexical association, including various likelihood measures and hypothesis tests. However, the experiments described in this paper use pointwise mutual information in order to retain consistency with SO-PMI-IR.

2.2 Semantic Spaces

While lexical association measures consider first-order similarity, semantic spaces measure second-order similarity. If words are similar in the second order they may not necessarily co-occur, but rather occur in similar contexts [14].

Semantic spaces represent concepts as a series of points in a large number of dimensions; the location of each point along each axis (or scale) is a measurement of the strength of association with that scale. Development of semantic spaces began in the field of cognitive science [28], where they were constructed by defining axes of interest and having several human subjects specify the position of each concept on that scale. For example, one might place the concepts of 'mouse' and 'mountain' at opposite ends of a scale that represents size. When populated in this way a semantic space can be conceptualised as a cuboid of data with dimensions of k concepts × m scales × n subjects.

The meaning of a concept (within the culture represented by the test subjects) can be represented by collapsing the cuboid along the subject dimensions for each concept [28]. $\mathbf{S}_{k,m}$ by finding the mean of each concept and scale combination over all subjects. This information can be augmented with a measure of variability in subject choices in order to evaluate how consistent the meaning of a given concept is within that culture. Furthermore, one can assess the similarity of concepts within that culture by applying a distance metric on the concept vectors extracted from the matrix.

Compiling a semantic space using human subjects is a laborious task, and furthermore it is subject to an arbitrary allocation of axes. Alternatively, a semantic space matrix may be constructed automatically from a corpus by passing a window over the corpus and counting the cooccurrences of features within that window [25]. The dimensions of the matrix then correspond to the features observed in the corpus. This can result in rather long vectors that represent the meaning of each word, but the dimensionality can be reduced, for example by disregarding infrequent word types [20], or by retaining only the most variant columns [6]. Semantic spaces constructed in this manner can be formalised as a quadruple, $\langle \mathtt{A}, \mathtt{B}, \mathtt{S}, \mathtt{M} \rangle$ [23]:

`B`: **Basis elements** `B` is a set of $b_{1..D}$ basis elements (where D is the number of dimensions in the space) and is analogous to the scales used in the cognitive science version of semantic spaces. `B` can be a set of document extracts [19], word types [25], word stems [24], dependency relations [29] or indeed any reasonable feature of a document. However, choosing basis elements is problematic because Zipf's Law states the vast majority of words appear infrequently; corpora may not yield reliable statistics for these words. It is necessary to accept a trade off between reliable estimations and breadth of coverage [23].

`A`: **Lexical association function** `A` is a lexical association function that maps co-occurrence frequencies of a target word, t, with basis elements so that w is represented as a vector $\mathbf{v} = [\mathtt{A}(b_1, w), \mathtt{A}(b_2, w), ..., \mathtt{A}(b_D, w)]$. This is akin to taking the mean of human-assigned scores for each concept in the cognitive science method. `A` is often simply the identity function or the reciprocal of the distance between t and b_i. However, this is unsatisfactory as raw cooccurrence counts can create a frequency bias [24]. The raw co-occurrence counts can be corrected for chance occurrences using measures such as Pointwise Mutual Information, the association function used in the experiments reported in this paper.

`S`: **Similarity measure** `S` is a similarity measure that maps pairs of vectors $\mathbf{v}$ and $\mathbf{w}$ onto a value that represents their contextual similarity. Applicable metrics include: Euclidean, City block, Cosine, Hellinger and Kullback-Leibler [21]. The experiments presented in this paper employed the Cosine measure as it maps to a value between -1 and 1, and mitigates any random scaling effects that might be caused by the range of the lexical association function and the number of basis elements [23]. The Cosine similarity measure is defined as:

$$\mathrm{sim}_{cos}(w, p) = 1 - \frac{\sum \mathbf{w}_b \mathbf{p}_b}{\sqrt{\sum \mathbf{w}_b{}^2}\sqrt{\sum \mathbf{p}_b{}^2}} \tag{3}$$

where $\mathbf{w}$ is the vector that represents word w and $\mathbf{p}$ is the vector that represents prototype p.

`M`: **Mapping transformation** `M` is a mapping of one semantic space onto another. A semantic space is fully functional without `M`, but the transformation can build a more structured model. One such mapping technique is Latent Semantic Analysis (LSA). LSA is based on the observation that text contains enormous quantities of weak interrelations that, if inference is applied correctly, can significantly expedite the process of language acquisition [19]. This supposition is represented mathematically by a dimensionality reduction of a semantic space; it is assumed that the reduced dimensions represent the latent interrelations. In LSA this is accomplished using the least-squares method of Singular Value Decomposition. A mapping transformation was not employed in the experiments described in this paper in order to retain comparability with other methods.

2.3 Distributional Similarity

Distributional similarity methods extend the notion of semantic spaces by considering the context of words as the set of grammatical relations featuring the word. The use of grammatical relations as context is potentially beneficial as it can reduce instances of cooccurrence which the window-based context used by the lexical association and semantic space methods would wrongly assume to be indicative of similarity [39]. Often the context taken is the set of all relations featuring the word, though recent research experimented with adding the notion of proximity from semantic spaces to distributional similarity calculations [29].

These methods are rooted in the distributional hypothesis [15], which suggests that there is a relationship between distributional and semantic similarity. This hypothesis has led to much research in automatic thesaurus construction [8, 14, 18, 22], estimating unseen feature frequency [9] and more recently to surveys of multiple techniques with application-based evaluations [39].

A frequently employed measure of distributional similarity is Lin's Measure [22], in which the similarity of two objects is the amount of information contained in the intersection of the objects' descriptions divided by the total information contained in the description of the objects. In the case of distributional similarity the information takes the form of frequencies of dependency triples, $||w, r, w'||$, which consist of two words (w and w') and the grammatical relation between them (r). The total amount of information, $T(w)$, described for any word w then is the frequency of all the dependency triples that match $||w, *, *||$ (where $*$ is a wild card). Lin's Measure is defined as:

$$\mathrm{sim}_{lin}(w, p) = \frac{\sum_{(r,w') \in T(w) \cap T(p)} I(w, r, w') + I(p, r, w')}{\sum_{(r,w') \in T(w)} I(w, r, w') + \sum_{(r,w') \in T(p)} I(p, r, w')} \tag{4}$$

The amount of information conveyed by a dependency relation is calculated using pointwise mutual information:

$$I(w, r, w') = \log \frac{||w, r, w|| \times ||*, r, *||}{||w, r, *|| \times ||*, r, w'||} \tag{5}$$

3. EXPERIMENTS

This section reports on optimisation and evaluation experiments conducted to assess Equation (1) and the word similarity methods when performing various tasks in sentiment analysis.

3.1 Experimental Setup

The following paragraphs describe the basic setup of the experiments that appear subsequently in this section.

3.1.1 Training Corpus

A key component of each of the techniques for determining word similarity described above is a source of word frequencies for probability estimates. Lexical Association methods are generally computationally inexpensive and so have been applied on very large corpora [37]. Experiments evaluating the performance of semantic spaces indicate that using as large a corpus as possible will yield better results [32], so in these experiments we sampled word occurrence and cooccurrence frequencies from the largest corpus available at the time this work was done, the Gigaword corpus [13]. The Gigaword corpus is a collection of newswire articles published by four international news agencies and contains around 1.7 billion words.

3.1.2 Feature selection

The following features (basis elements in the Semantic Spaces terminology) were investigated for use by the lexical association and semantic space methods: lemmatised words, lemmatised words with part-of-speech tags, and adjectives and adverbs only. The basis elements of the distributional similarity method are the grammatical relations acquired using the second release of the RASP dependency parser [5].

The effect of varying the number of features was also investigated. Previous experimental results showed that ordering the features by frequency and using as many as computationally feasible produces the best results [24]. The number of features is a very important parameter for practical purposes, however, as it greatly effects runtime. It was therefore useful to evaluate whether gains may be made from constraining the number of features in this application, so we considered a logarithmic scale of numbers of features: 5000, 10000, 20000, 40000, 80000, 160000 and 320000.

As PMI is unreliable when applied to low-frequency words [7, 37] it was not calculated for words occurring less than four times in the corpus. This constraint was applied to each similarity measure in order to maintain comparability between the techniques.

3.1.3 Prototype Selection

The basic method requires a set of prototypical examples to which problem words may be compared. SO-PMI-IR employs just one prototype for each class, but more recent experimental results suggest that providing several prototypes is beneficial [42].

Ideally, a set of prototypes should broadly represent a class, whilst each individual prototype should be unambiguous with respect to the class. Our approach to collecting prototypical words began by obtaining the synonyms of class labels from Roget's Thesaurus. Words obtained from Roget's Thesaurus are appealing since while not strictly being synonyms (as entries contain various parts-of-speech for a given concept), they are highly semantically-related and can increase the breadth of context observed. We then grew the set of prototypes using from entries in WordNet synsets [11] and inspected the senses of each word, disregarding any that did not unambiguously represent the class[1]. Finally, as using many prototypes is computationally expensive when employing semantic spaces or distributional similarity, we limited each set to the seven most frequent examples in the Gigaword corpus. The resulting prototypes are listed in Table 1.

Class	Prototypes
Positive	benefit, best, excellent, good, nice, perfect, supreme
Negative	abuse, bad, disastrous, evil, outrage, sad, wrong

Table 1: Polarity prototypes obtained from Roget's Thesaurus and WordNet, and selected based on frequency in the Gigaword corpus.

3.1.4 Measuring Performance

The performance of the techniques is measured in terms of precision (P, the proportion of correct identifications relative to the total number of identifications made), recall (R, the proportion of correct identifications relative to the total number of possible identifications), and F-measure (F_1, the harmonic mean of the precision and recall). The reported evaluations of statistical significance of experimental results utilised paired t-tests.

3.2 Constructing a Polarity Lexicon

To investigate the capabilities of the word similarity methods in determining the polarity of individual words we conducted an experiment in which entries from the General Inquirer (GI) [35] were classified as positive or negative. After disregarding words with multiple senses marked with conflicting polarities, the GI contains 1,374 positive words and 1,708 negative words. A random third of these words were used for optimisation purposes while the rest formed a test data set.

The results of applying the experimental setup described above indicated that each word similarity method performed best when using plain lemmas, and when using the maximum number of basis elements. The semantic space method was best with a cooccurrence window of 3 words, while the lexical association method was best with a window of 10 words.

Table 2 lists the results of evaluating the optimal parameters for each similarity method on the GI test set. It also includes three baselines: labelling all entries as the majority class (negative), labelling entries randomly, and labelling entries if and only if they were prototypes. Semantic spaces performed better than the lexical association and distributional similarity methods, which achieved similar results. All three word similarity methods performed markedly better than the baselines, and all differences were significant.

3.3 Scoring Sentences According to Strength of Sentiment

This section reports experiments that evaluate the efficacy of the word similarity methods in a reproduction of SemEval 2007's shared task on Affective Text [36], which involved scoring sentences according to their strength of sentiment and six emotion types[2].

[1] To assess the difficulty of selecting which WordNet senses are pertinent to a particular class, we asked two annotators to independently label WordNet sense glosses as being relevant or irrelevant to one of six Basic emotions [27]. The experiment included 100 words and 464 gloss definitions; the annotators agreed on the relevance/irelevance of 85% of glosses ($\kappa = 0.64$). Reframing the task to determine whether annotators agreed that a word unambiguously represented a class (i.e. all of its senses were relevant), we found that the annotators agreed on 90% of words ($\kappa = 0.77$).

[2] A useful feature of the word similarity methods is that, given an appropriate set of prototypes, they are readily transferred to other classification problems such as this. The

	F_1	P	R
Semantic Space	0.816	0.838	0.796
Lexical Association	0.657	0.717	0.607
Distributional Similarity	0.643	0.676	0.612
Majority	0.554	0.554	0.554
Random	0.500	0.500	0.500
Prototypes	0.006	1.000	0.003

Table 2: The performance of the word similarity methods in classifying General Inquirer entries according to polarity.

	F_1	P	R	r
Semantic Space	0.020	0.444	0.010	0.502
Distributional Similarity	0.302	0.531	0.211	0.466
Lexical Association	0.160	0.464	0.097	0.406
CLaC	0.160	0.614	0.092	0.477
UPAR7	0.153	0.575	0.088	0.370
SWAT	0.063	0.457	0.034	0.353
CLaC-NB	0.425	0.312	0.664	0.254
SICS	0.386	0.284	0.602	0.207

Table 3: The performance of the word similarity methods in the Valence sub-task of the Affect Text task, compared with original participants.

When assessing the word similarity methods described in this chapter on the Affective Task it was necessary to translate the scores calculated into the range prescribed by the task designers (-100 represents a highly negative sentence and +100 indicates a highly positive sentence). To accomplish this each sentence was tokenised into a vector of words ($\mathbf{W}$) and scored (refer to Equation 1) as:

$$\text{valence}\left(\mathbf{W}\right) = \text{score}\left(\mathbf{W}, p\right) - \text{score}\left(\mathbf{W}, n\right) \qquad (6)$$

where p is a set of positive prototypes and n is a set of negative prototypes. The maximum absolute valence score over all sentences was then used as a normalising constant. The resulting scores were evaluated using Pearson's correlation coefficient (r), and according to precision, recall and F-measure by mapping scores to negative if less than -50, positive if greater than 50, and neutral otherwise.

Using development data provided for the Affective Text task and the experimental setup described above, we found that both the lexical association and semantic space methods performed best when considering just 20,000 lemmas over a cooccurrence window of 4 features. The distributional similarity method was best when using 160,000 lemmas.

Table 3 lists the results of the optimised word similarity methods applied to the test data of the Affective Text valence sub-task. The table also lists the results of the original participants of the sub-task, which employed a range of techniques including: supervised machine learning (CLaC-NB, SWAT), weakly-supervised learning (SICS) and knowledge-based approaches (UPAR7, CLaC).

The semantic space method performs better than any other system on the fine-grained valence task, with a correlation score of 0.502. This does not translate to a good performance on the coarse evaluation though, where a system using supervised machine learning performs best. The coarse-grained evaluation of the word similarity methods could perhaps be improved by applying a more sophisticated transformation from scores to the desired range.

	F_1	P	R
Lexical Association	0.687	0.687	0.687
Semantic Space	0.667	0.667	0.667
Distributional Similarity	0.608	0.608	0.608
Random	0.500	0.500	0.500
Prototypes	0.492	0.635	0.401

Table 4: The performance of weakly supervised methods in determining the sentiment of movie reviews.

3.4 Sentiment Classification of Movie Reviews

We next assessed the capabilities of the word similarity techniques at the document level in a task that involves the classification of movie reviews in the Polarity 2.0 data set [30] as positive or negative. One third of the reviews was taken for development data, with the rest reserved for testing data. Optimising each word similarity method on the development data, we found that the maximum number of plain lemmas (320,000) was the optimal feature setup for all methods. Lexical association performed best with a cooccurrence window of 8 words and semantic space was best with 3 words.

Table 4 lists the results of evaluating the optimally parameterised similarity methods on the testing data, and compares the results to two baselines: choosing positive or negative at random, and counting the frequency of prototypes in reviews. With an F_1 score of 0.687, the lexical association methods slightly outperforms the semantic space method, though this difference is not significant. Both the lexical association and semantic space methods significantly outperform the distributional similarity method, and the performance of each word similarity method is significantly better than the baselines.

Being only weakly-supervised, these methods are less effective than supervised techniques complemented with subjectivity detection (Support Vector Machines achieved an accuracy of 86.2% on this dataset [30]). There appears to be no previous work published on weakly-supervised methods applied to the Polarity 2.0 data. However, SO-PMI-IR was applied to a set of movie reviews, achieving an accuracy of 65.8% [38]. While it is not appropriate to strictly compare these results, being obtained from different data sets, it is nevertheless included here to give an indication of the difference in performance of the methods.

4. PERFORMANCE ACROSS TOPICS, DOMAINS AND TIME-PERIODS

One motivation for investigating the word similarity methods described in Section 2 is that they are trained on a large quantity of general text, and so performance is less likely to vary across topics, domains and time-periods. To assess the techniques' usefulness in this respect, we reproduced previous experiments [33] that investigated dependencies on training data in supervised techniques.

results on the emotion sub-task are not reported here, however, due to space constraints. Please see [34] for details.

	FIN	M&A	MIX	Mean
NB-FIN	0.803	0.755	0.740	0.765
NB-M&A	0.775	0.753	0.758	0.762
NB-MIX	0.707	0.629	0.846	0.717
SVM-FIN	0.788	0.727	0.689	0.732
SVM-M&A	0.745	0.755	0.755	0.752
SVM-MIX	0.720	0.689	0.811	0.737
Lexical Association	0.711	0.704	0.708	0.708
Semantic Space	0.714	0.710	0.709	0.711
Distribution Similarity	0.685	0.678	0.683	0.682

Table 5: The accuracies of supervised and weakly-supervised methods in classifying newswire articles (FIN=Finance, M&A=Mergers & Acquisitions, MIX=a discrete mix of both types) according to sentiment in various topics of financial news, with the harmonic means of the accuracies.

	News	Pol. 1.0	Mean
NB-Newswire	0.782	0.576	0.663
NB-Polarity 1.0	0.532	0.789	0.636
SVM-Newswire	0.782	0.632	0.699
SVM-Polarity 1.0	0.636	0.815	0.714
Lexical Association	0.708	0.687	0.697
Semantic Space	0.711	0.667	0.688
Distributional Similarity	0.682	0.645	0.663

Table 6: The accuracies of supervised and weakly-supervised methods in classifying documents in the domains of newswire articles and movie reviews, with the harmonic means of the accuracies.

Table 5 shows the performance of the various methods when applied to newswire articles of different topics, Table 6 lists their performance across the domains of newswire articles and movie reviews, while Table 7 contains the performance across sets of movie reviews from different time periods. Note that, in each of these tables the first set of results are repetitions of supervised machine-learning results [33] (NB=Naïve Bayes, SVM=Support Vector Machines, while the suffix refers to the data set on which the model was trained). The second set of results are that of the word similarity methods described in Section 2 (Lexical Association, Semantic Spaces and Distributional Similarity).

The word similarity methods give consistent results across domain, topic and time-period (a paired t-test found that the differences between sets in each of the three experiments are not significant). However, the results show that, with respect to topic and temporal dependency it would be more effective to use the supervised techniques (in particular Support Vector Machines) and accept some loss in performance when processing data of different topic or time-period than to use the word similarity methods. In the domain dependency experiments, though, the differences are small, which suggests that using methods like lexical association and semantic spaces may be appropriate when the task involves determining the sentiment of documents from various domains.

	Pol. 1.0	Pol. 2004	Mean
NB-Polarity 1.0	0.789	0.718	0.752
NB-Polarity 2004	0.632	0.765	0.692
SVM-Polarity 1.0	0.815	0.775	0.794
SVM-Polarity 2004	0.765	0.808	0.786
Lexical Association	0.687	0.691	0.689
Semantic Space	0.667	0.679	0.673
Distributional Similarity	0.606	0.606	0.606

Table 7: The accuracies of supervised and weakly-supervised methods in classifying movie reviews from data sets representing different time-periods, with the harmonic means of the accuracies.

5. RELATED WORK

Some alternative approaches to resolving dependency in supervised sentiment classification are: to train classifiers on a collection of topics, domains or time-periods; to create a voting ensemble of classifiers trained on different topics, domains, or time-periods; or to combine labelled data with unlabelled data representing the target data set [3]. One might also consider selecting features that are independent of topic, domain or time-period [10], or supplementing a classifier with a small number of labelled examples (50) from the target domain [2].

Other researchers investigated the structural corresponding learning (SCL) algorithm for the problem of adapting supervised sentiment classifiers to new domains [4]. Using labelled data from a source domain and unlabelled data from the source and target domains, SCL uses the correlation of pivot features with all other features in order to learn which non-pivots predict pivots. When adapting the source classifier to the target domain, the pivot-predicting features of the source domain are projected onto the pivot-predicting features of the target domain. For instance, in reviews of computers *fast dual-core* may predict *excellent*, while in a mobile phone review *good quality reception* might also predict *excellent*. When SCL adapts a classifier from the computer review domain to the mobile-phone domain, the feature *fast dual-core* would project on to the feature *good quality reception.*

As mentioned previously, the performance of weakly supervised techniques for sentiment classification is less likely to vary across domains, topics and time periods. Another weakly-supervised technique is to employ iterative retraining [43], where in each repetition the relative frequency of prototypes in zoned text is inspected, and thus classified as positive or negative to create two sub-corpora. New prototypes are then obtained from the relative frequency of words in each sub-corpus, and the process is repeated. Alternatives include knowledge-based approaches using resources such as Word-Net, for example by calculating the minimum path length from a target word to a prototype [16] or by utilising fuzzy sets [1].

6. CONCLUSIONS

This paper proposed that word-similarity techniques for sentiment analysis can operate independently of domain, topic and time-period, as they are only weakly-supervised, requiring just a few prototypical examples of sentiment and training on a very large corpus of general text. It described three techniques for measuring word similarity: lexical association, semantic spaces and distributional similarity. Investigating the capabilities of the word similarity methods in performing various tasks in sentiment analysis, we found that the semantic space method performed well when compiling a polarity lexicon and scoring sentences according to their strength of sentiment. Both the lexical association and semantic space methods performed well when classifying movie reviews. All the methods significantly improved on baselines in each of the three tasks.

Being only weakly-supervised, these methods are less effective than supervised techniques. The weakly-supervised methods may still be of practical benefit, however, if the methods' performance does not vary greatly across different topics, domains and time-periods as proposed. We therefore repeated previous topic, domain and time-period dependency experiments [33], finding that the word similarity techniques give reasonably consistent results across these data sets. However, the results showed that, with regards to topic and temporal dependency, it is still more effective to use supervised machine learning and accept some loss in performance when analysing data of a different topic or time-period. The differences between the machine learning and the word similarity methods were small, indicating that the word similarity methods may be appropriate when the task involves data from a variety of domains.

Future work will investigate further gains in the performance of the word similarity methods may be gained by: automatically removing objective sentences in the training and testing data using subjective language detection [41]; tailoring the training data to the target domain using keywords; experimenting with feature selection; and growing the set of prototypes using bootstrapping by selecting the highest scoring candidates in each iteration.

7. ACKNOWLEDGMENTS

We would like to thank Roy Lipski for the annotated newswire articles and Bo Pang for the annotated movie reviews. We would also like to thank the UK EPSRC, who supported the work of the first author with a studentship. Thanks also to the anonymous reviews for their helpful feedback.

8. REFERENCES

[1] A. Andreevskaia and S. Bergler. 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*, Trento, Italy, 2006.

[2] A. Andreevskaia and S. Bergler. When specialists and generalists work together: Overcoming domain dependence in sentiment tagging. In *Proceedings of the Joint 46th Annual Meeting of the Association for Computational Linguistics (ACL-2007) and Human Language Technology Conference (ACL08: HLT)*, pages 290–298, Columbus, Ohio, USA, June 2008.

[3] A. Aue and M. Gamon. Customizing sentiment classifiers to new domains: A case study. In *Proceedings of Recent Advances in Natural Language Processing (RANLP)*, 2005.

[4] J. Blitzer, M. Dredze, and F. Pereira. Biographies, Bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL-2007)*, pages 440–447, Prague, Czech Republic, June 2007.

[5] T. Briscoe, J. Carroll, and R. Watson. The second release of the rasp system. In *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*, pages 77–80, Sydney, 2006.

[6] C. Burgess and K. Lund. Representing abstract words and emotional connotation in a high-dimensional memory space. In *Proceedings of the Cognitive Science Society*, pages 61–66, Hillsdale, N.J., 1997. Lawrence Erlbaum Associates, Inc.

[7] K. W. Church and P. Hanks. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16:22–29, 1990.

[8] J. R. Curran and M. Moens. Improvements in automatic thesaurus extraction. In *Proceedings of the ACL-SIGLEX Workshop on Unsupervised Lexical Acquisition*, pages 59–67, Philadelphia, PA, USA, 2002.

[9] I. Dagan, L. Lee, and F. Pereira. Similarity-based models of word cooccurrence probabilities. *Machine Learning*, 34(1–3):43–69, 1999.

[10] C. Engström. Topic dependence in sentiment classification. M. Phil., St Edmund's College, University of Cambridge, 2004.

[11] C. Fellbaum, editor. *Wordnet: An Electronic Lexical Database.* The MIT Press, Cambridge, MA, 1998.

[12] M. Gamon. Sentiment classification on customer feedback data: Noisy data, large feature vectors, and the role of linguistic analysis. In *Proceedings of the 20th International Conference on Computational Linguistics*, Geneva, Switzerland, 2004.

[13] D. Graff. English gigaword. Linguistic Data Consortium, Philadelphia, 2003.

[14] G. Grefenstette. Corpus-derived first-, second- and third-order word affinities. In *Proceedings of Euralex*, pages 279–290, Amsterdam, 1994.

[15] Z. S. Harris. *Mathematical structures of language.* John Wiley, New York, 1968.

[16] J. Kamps, M. Marx, R. J. Mooken, and M. de Rijke. Using WordNet to measure semantic orientations of adjectives. In *Proceedings of the Language Resource and Evaluation Conference*, 2004.

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

[18] A. Kilgarriff. Thesauruses for natural language processing. In *Proceedings of the Joint Conference on Natural Language Processing and Knowledge Engineering*, pages 5–13, Beijing, China, 2003.

[19] T. K. Landauer and S. Dumais. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104(2):211–240, 1997.

[20] J. P. Levy and J. A. Bullinaria. Learning lexical properties from word usage patterns: Which context

words should be used? In R. M. French and J. P. Sougne, editors, *Connectionist Models of Learning, Development and Evolution: Proceedings of the Sixth Neural Computation and Psychology Workshop*, pages 273–282, London, 2001. Springer.

[21] J. P. Levy, J. A. Bullinaria, and M. Patel. Explorations in the derivation of word co-occurrence statistics. *South Pacific Journal of Psychology*, 10(1):99–111, 1998.

[22] D. Lin. Automatic retrieval and clustering of similar words. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, 1998.

[23] W. Lowe. Towards a theory of semantic space. In *Proceedings of the 6th Neural Computation and Psychology Workshop*, pages 303–311. Springer Verlag, 2001.

[24] W. Lowe and S. McDonald. The direct route: Mediated priming in semantic space. In *Proceedings of the 22nd Annual Meeting of the Cognitive Science Society*, 2000.

[25] K. Lund and C. Burgess. Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments & Computers*, 28(2):203–208, 1996.

[26] T. Mullen and N. Collier. Sentiment analysis using support vector machines with diverse information sources. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, Barcelona, Spain, 2004.

[27] A. Ortony and T. J. Turner. What's basic about basic emotions? *Psychological Review*, 97(3):315–331, 1990.

[28] C. Osgood, G. Suci, and P. Tannenbaum. *The measurement of meaning*. University of Illinois Press, Urbana, U.S.A., 1957.

[29] S. Padó and M. Lapata. Dependency-based construction of semantic space models. *Computational Linguistics*, 33(2):161–199, 2007.

[30] B. Pang and L. Lee. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, Barcelona, Spain, 2004.

[31] B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up? Sentiment classification using machine learning techniques. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing*, Philadelphia, PA, USA, 2002.

[32] M. Patel, J. A. Bullinaria, and J. P. Levy. Extracting semantic representations from large text corpora. In *Proceedings of the Fourth Neural Computation and Psychology Workshop*, pages 199–212, London, 1997. Springer-Verlag.

[33] J. Read. Using emoticons to reduce dependency in machine learning techniques for sentiment classification. In *Proceedings of the ACL Student Research Workshop*, Ann Arbor, MI, USA, 2005.

[34] J. Read. *Weakly Supervised Techniques for the Analysis of Evaluation in Text*. PhD thesis, University of Sussex, forthcoming.

[35] P. J. Stone. *The General Inquirer: A computer approach to content analysis*. The MIT Press, 1966.

[36] C. Strapparava and R. Mihalcea. SemEval-2007 task 14: Affective text. In *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval 2007)*, Prague, Czech Republic, June 2007.

[37] P. D. Turney. Mining the Web for synonyms: PMI-IR versus LSA on TOEFL. In *Proceedings of the 12th European Conference on Machine Learning*, Berlin, 2001.

[38] P. D. Turney. 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, PA, USA, 2002.

[39] J. Weeds and D. Weir. Co-occurrence retrieval: A flexible framework for lexical distributional similarity. *Computational Linguistics*, 31(4):439–475, 2005.

[40] J. Wiebe, E. Breck, C. Buckley, C. Cardie, P. Davis, B. Fraser, D. Litman, D. Pierce, E. Riloff, T. Wilson, D. Day, and M. Maybury. Recognizing and organizing opinions expressed in the world press. In *Proceedings of the AAAI Spring Symposium on New Directions in Question Answering*, 2003.

[41] J. Wiebe, T. Wilson, R. Bruce, M. Bell, and M. Martin. Learning subjective language. *Computational Linguistics*, 30(3):277–308, 2004.

[42] H. Yu and V. Hatzivassiloglou. Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, Sapporo, Japan, 2003.

[43] T. Zagibalov and J. Carroll. Unsupervised classification of sentiment and objectivity in Chinese text. In *Proceedings of the Third Joint International Conferences on Natural Language Processing (IJCNLP)*, Hyderabad, India, 2008.