Topic-Based Sentiment Analysis for the Social Web: The role of Mood and Issue-Related Words[1]

Mike Thelwall, Kevan Buckley
Statistical Cybermetrics Research Group, School of Technology, University of Wolverhampton, Wulfruna Street, Wolverhampton WV1 1SB, UK.
E-mail: m.thelwall@wlv.ac.uk, K.A.Buckley@wlv.ac.uk
Tel: +44 1902 321470 Fax: +44 1902 321478

General sentiment analysis for the social web has become increasingly useful to shed light on the role of emotion in online communication and offline events in both academic research and data journalism. Nevertheless, existing general purpose social web sentiment analysis algorithms may not be optimal for texts focussed around specific topics. This article introduces two new methods, mood setting and lexicon extension, to improve the accuracy of topic-specific lexical sentiment strength detection for the social web. Mood setting allows the topic mood to determine the default polarity for ostensibly neutral expressive text. Topic-specific lexicon extension involves adding topic-specific words to the default general sentiment lexicon. Experiments with eight data sets show that both methods can improve sentiment analysis performance in corpora and are recommended when the topic focus is tightest.

Introduction

The rise of the social web has created new opportunities to track reactions to events via public texts, such as tweets. An important aspect of public opinion and reactions is sentiment: whether people feel positive or negative towards an event and how this changes over time. Studies combining the easy availability of social web texts and other data with general sentiment analysis algorithms have been able to give new insights into human behaviour as a result (e.g., Chung & Mustafaraj, 2011; Dodds & Danforth, 2010; Gruzd, Doiron, & Mai, 2011; Kramer, 2010) and this approach shows promise for data journalism (e.g., The Guardian, 2012). This social science use of sentiment analysis is different from its most common use for detecting the polarity of opinions of consumer products or movies (Pang & Lee, 2008; Liu, 2012) although it can exploit similar methods.

An important problem with many types of sentiment analysis is that it can be highly topic dependent in the sense that algorithms that predict sentiment accurately on texts from one specific domain may be much less accurate on another (Aue & Gamon, 2005; Tan, Wu, Tang, & Cheng, 2007). Thus, whilst general sentiment analysis algorithms may give reasonable overall performance, it seems likely that there will be scope for improving them by customising them for specific topics. For some narrow topics, such customisation may be required to give reasonable performance if sentiment is routinely expressed with a specialist or rare vocabulary (e.g., *arrest* and *burn* may be key sentiment indicators in riot discussions). Hence, methods are needed to either develop topic-specific sentiment analysis algorithms for the social web or to customise general social web sentiment analysis algorithms for specific domains.

[1] This is a preprint of an article to be published in the Journal of the American Society for Information Science and Technology

As argued below, most sentiment analysis algorithms are not suitable for many social sciences goals because they can harness non-sentiment features of text that associate with sentiment, such as politicians' names. This can give more accurate results overall if negative sentiment is expressed in obscure ways, such as with sarcasm and irony, so that the politician's name (or other non-sentiment feature) is the best clue for the presence of negativity. Nevertheless, this sacrifices the face validity of the results and hence their utility for social science because it can falsely identify sentiment in a systematic way. For example, a set of positive or neutral texts about a generally disliked politician, such as a holiday report, could appear as a burst of negativity because of the presence of the politician's name in the texts (see also the example in the conclusion section below). To avoid this issue, sentiment analysis can rely on predefined sets of sentiment-related terms and ignore non-sentiment features (i.e., a lexical approach). Whilst significant research within the field of sentiment analysis of product reviews (often called opinion mining) has been devoted to the domain transfer problem, most does not use lexical sentiment analysis and none seems to be directly relevant for social web data sets. As discussed below, domain transfer research typically involves taking an opinion mining algorithm designed for one review domain (e.g., movies) and customising it for a different domain (e.g., digital cameras), for example by mapping sentiment features from one domain onto apparently similar features in the other. It is not clear whether such methods will also work for general sentiment analysis algorithms.

This article introduces two new methods to improve lexical sentiment analysis for the social web by allowing a general algorithm to be customised for a specific topic. The first method takes into account the mood of the posts within the topic and the second method identifies and adds topic-specific terms to the general sentiment lexicon. The methods are evaluated using an existing general lexical sentiment analysis algorithm for the social web, SentiStrength (Thelwall, Buckley, & Paltoglou, 2012; Thelwall, Buckley, Paltoglou, Cai, & Kappas, 2010).

Mood setting and lexicon extension

The sentiment strength detection software SentiStrength is designed for social web texts and predicts the strength of both positive and negative sentiment in texts simultaneously using a lexical approach (Thelwall et al., 2012; Thelwall et al., 2010). SentiStrength assigns a positive sentiment strength of 1 (no positive sentiment) to 5 (very strong positive sentiment) *and* a negative sentiment strength of -1 (no negative sentiment) to -5 (very strong negative sentiment) to each text (zeros are not used). The core of SentiStrength is a list of 2,608 words and word stems together with a typical positive or negative sentiment strength for social web texts. For example, *good* scores 3 and *scare[…]* scores -4 (matching scare, scared etc.) and so the sentence "I was scared but it was good" might score +3 on the positive scale *and* -4 on the negative scale. If there are multiple sentiment words then the strongest sentiment word is chosen and if there are no sentiment words then a neutral sentiment is assumed. In addition, there are special rules for dealing with negations, questions, booster words (e.g., very), emoticons, and a range of other special cases (Thelwall et al., 2012; Thelwall et al., 2010). If a single overall number is needed for sentiment strength then the positive number can be added to the negative number to give a score in the range -4, -3,… 4. SentiStrength runs in two modes: supervised and unsupervised. In unsupervised mode, it uses the pre-defined sentiment strength weights for the lexicon. In supervised mode, it uses a training set of data to automatically adjust the

lexicon term weights to give more accurate results. For instance, if the term *scared* was generally not used in a negative context in the corpus (e.g., of horror movie discussions) then this process might leave it with a neutral score of -1 (or even a positive score). Previous tests have shown both supervised and unsupervised versions to have similar levels of accuracy on a range of social web texts (Thelwall et al., 2012).

An extension of SentiStrength for the current paper is the option to output a single scale classification from -4 (strongly negative overall) to +4 (strongly positive overall). This is essentially the same scheme as used in SO-CAL (Taboada, Brooke, Tofiloski, Voll, & Stede, 2011).

Mood setting One of the rules in SentiStrength is that two indicators of "energy" that are not associated with a recognised sentiment term are assumed to be positive. These indicators are punctuation at the end of sentences that includes exclamation marks and repeated letters added to the spelling of a word (e.g., Miiiiike) in a way that seems to be used for emphasis or to convey enthusiasm in computer mediated communication. With the social network comments used to create the initial versions of SentiStrength, human coders tend to interpret energy as positive if there was no associated evidence of a negative polarity (Thelwall et al., 2010). The new mood rule is to allow this to be changed to negative as the default setting for a specific topic. The rationale for this is that if the mood of a topic is negative then the reader may use this mood to infer the likely sentiment of an expression of sentiment or energy that is otherwise apparently neutral.

Lexicon extension As discovered in domain-specific opinion mining research for product reviews (see below), individual words may typically be positive when associated with one topic but negative when associated with another. For example, "large" may tend to be positive for reviews of televisions but negative for reviews of electronic cameras. The social web approach of SentiStrength and similar programs relies upon a large lexicon of words with a pre-defined average sentiment. Nevertheless, for a specific topic there may be rare words or specialist words that are frequently used to express sentiment. The lexical extension method is an attempt to identify these words and use them to improve sentiment strength prediction through a topic-specific lexicon extension – a set of words and word strengths. This is a supervised method and requires a sentiment annotated set of texts from a specific topic.

The lexical extension method is as follows. For each term in each text in the training set, a tally is kept of the difference between the human coded value of texts containing the term and their SentiStrength classifications. For example, suppose that the term *brown* occurs in text A with human codes (3,-1) and text B with human codes (4, -2) and SentiStrength gives classifications A (2, -2) and B (3, -1). Then for positive sentiment the difference would be (3-2) + (4-3) = 2 and for negative sentiment the result is (-1 - -2) + (-2 - -1) = 0. This suggests that the term brown may express positive sentiment for this topic and that adding it to the lexicon with a positive strength may improve the performance of the classifier. The extension method is not fully automatic, however, because it may identify nouns, such as politicians' names, that are typically used in a positive or negative context without expressing sentiment. This can be undesirable in some contexts (see below). The large number of terms assessed is statistically likely to produce some random irrelevant terms and systematic biases in the data can result in common terms having high scores. In both cases the terms may reduce performance, if added. Hence the main method is to automatically produce a list of the *n* terms that have the greatest total positive or negative total distance values and a list of the *n* terms that have the greatest total positive or

negative average distance values for a human coder to select the terms that can legitimately be added. This gives a total of up to 4*n* terms to be checked for the scale sentiment classification method and 8*n* for the dual sentiment classification method (4*n* for positive and 4*n* for negative).

A secondary lexical method is as above but with some automatic checking of the human coder results. For this, each individual term selected by the human coder is automatically checked on a human coded test corpus to see whether it improves performance overall, rejecting those that do not. This stage may be useful if a human coder selects terms that are apparently valuable but in practice are not useful due to polysemy or other reasons.

Literature review

This review discusses common sentiment analysis approaches and the face validity problems that they may generate for a sentiment analysis of social web topics. For a literature review of more general issues related to the overall design of SentiStrength see the two previous papers (Thelwall et al., 2012; Thelwall et al., 2010) or more general reviews (Pang & Lee, 2008; Liu, 2012). For example, the review does not cover sub-document level sentiment analysis (e.g., Wilson, 2008).

There are two common general approaches to sentiment analysis, machine learning and lexical, although many algorithms have elements of both. The machine learning approach usually starts with a simple sentiment rule (i.e., unsupervised machine learning) or a collection of texts that have been annotated by human coders for sentiment (i.e., supervised machine learning) and then learns text features that associate with positive or negative sentiment (Mejova, & Srinivasan, 2011; Pang, Lee, & Vaithyanathan, 2002; Riloff, Patwardhan, & Wiebe, 2006; Zagibalov, 2010). These features are typically unigrams, bigrams and trigrams: sets of one to three consecutive words in texts that normally associate with sentiment. For instance, the sentiment features identified may include: *like*, *shocked_to_discover*, and *george_bush*. These features are then used to predict the sentiment of new texts. A significant disadvantage of this approach for social science uses is that it can extract non-sentiment features that associate with sentiment because they are frequently used in sentences of a particular polarity. For instance, politicians' names and country names may typically be associated with strong positive or negative feelings within discussions on a specific topic (e.g., the Middle East). Hence, if a machine learning approach is used to identify patterns in sentiment then it is likely to identify patterns that are partly related to discussions of entities that are not directly sentiment related (e.g., *hamas_will*, *George_Galloway*, *Israel_will*, and *that_Palestinians* in Thelwall et al., 2012). For example, a burst of negativity in Twitter may be "detected" when an unpopular president gives a speech on a neutral or positive topic simply because his or her name is mentioned often in the context of the speech.

The lexical approach starts with a collection of terms with known sentiment associations and then applies a set of rules to predict the sentiment of texts based upon the occurrence of these words. These sentiment word lists have been taken from various pre-existing sources, such as the General Inquirer lexicon (Stone, Dunphy, Smith, & Ogilvie, 1966), the LIWC program (Pennebaker, Mehl, & Niederhoffer, 2003) or WordNet (Agerri & García-Serrano, 2010; Baccianella, Esuli, & Sebastiani, 2010; Strapparava, Valitutti, & Stock, 2006). Lexicons have also been built semi-automatically, starting with a seed set of sentiment terms (Taboada, Anthony, & Voll, 2006; Turney, 2002) or manually from a

development corpus (Taboada et al., 2011). General lexicons can also be built semi-automatically from a general corpus by heuristics such as starting with a set of sentiment words of known polarity and then identifying new terms that are frequently connected to them by "and", implying that the words have the same polarity, or terms such as "but" that imply an opposite polarity (Hatzivassiloglou, & McKeown, 1997; see also Kaji & Kitsuregawa, 2007). The rules applied to sentiment words to predict sentiment deal with aspects of language such as negation and the use of boosters/intensifiers (e.g., very) (Taboada et al., 2011) and sometimes requires a linguistic parsing of text for parts-of-speech. A limitation of generic lexical methods for the analysis of specific social web topics is that they can assign incorrect sentiments on the basis of sentiment terms that are frequently used in the topic but with a different sentiment to that of their normal usage. For instance, the terms *dip* and *pistol* might have negative associations in normal usage but be neutral or positive in discussions of the Olympics.

Aspect-based sentiment analysis is a special type of sentiment analysis that identifies the sentiment orientation of a text towards the different product aspects discussed. Aspect lexicons can include terms that have a polarity relative to a specific aspect of the opinion target, such as *large* being positive for a phone screen but negative for the phone overall (e.g., Ding, Liu, & Yu, 2008; Hu & Liu, 2004; Lu, Castellanos, Dayal, & Zhai, 2011). It is also possible to learn the implied sentiment of expressions in a particular context composed entirely of non-sentiment terms (Zhang, & Liu, 2011). Whilst aspect-based sentiment analysis methods are capable of detecting sentiment on a fine-grained level and could, therefore, avoid errors caused by interpreting terms out of context, they are not designed to detect the overall sentiment of a text, just that of relevant objects within it. Aspect-based sentiment classifiers are also still capable of misidentifying non-sentiment terms that are commonly used in a negative (positive) context because they appear statistically to be negative (positive). For a specific social web topic, such terms could include politicians' names (causing the politicians to always be cast as negative or positive) and words commonly used in descriptions of key events (causing the events to be always cast as negative or positive).

Domain-specific sentiment classifiers can be more accurate than general purpose sentiment classifiers. For example, in a review, non-sentiment terms may strongly associate with sentiment (e.g., *4G* may be positive and *heavy* may be negative in phone reviews) and this may be the key to good performance. The same terms may be irrelevant to other types of reviews, however, such as for movies or washing machines, and so an opinion mining algorithm designed for one domain may fail or perform poorly on another (Aue & Gamon, 2005). To avoid building classifiers from scratch for a new domain, the results of an ensemble of different domain-specific classifiers can be combined to predict the sentiment of the new texts (Aue & Gamon, 2005). Alternatively, ways of expressing sentiment in domains with existing classifiers could be matched with similar expressions in a new domain without a classifier (Blitzer, McDonald, & Pereira, 2006). A related method is to construct a graph based upon inter-document similarity, mixing the texts from the two domains and using graph algorithms to help predict the sentiment of texts in the new domain (Wu, Tan, & Cheng, 2009). Other researchers have also focused on developing effective methods of representing documents in this context (Glorot, Bordes, & Bengio, 2011) and of identifying relevant domains for effective domain transfer (Ponomareva & Thelwall, 2012). Domain transfer methods inherit the face validity problems for social web topics of the classifiers that they are based upon, however, and seem likely to reduce the face validity of the results

of a sentiment classification still further because they are all based upon an assumption of some type of similarity between the source and target domains.

Research Questions

The research questions address the issue of whether the two methods described above can improve the performance on specific topics. The rules are not designed for general purpose fully-automated sentiment analysis but for sentiment strength detection on specific topics of social science interest.

1. Can a negative mood assumption improve sentiment strength detection in the social web for any topics?
2. Can lexicon extensions improve sentiment strength detection in the social web for any topics?

The research questions were addressed by implementing the features within SentiStrength and then evaluating them on suitable test data.

Methods

Two main topics were selected to test the two new methods. Both are specific social sciences issues that were investigated through social web data from Twitter. These were selected as data arising from real-world applications of sentiment analysis. Both data sets were selected by social scientists at the UK National Centre for Text Mining in Manchester as being of intrinsic social science interest and supplied to the SentiStrength creators in Wolverhampton to run a sentiment analysis. Hence the data sets are both genuine examples of contexts in which a social science sentiment analysis was required by an external client[2].

The *UK riot rumours corpus* is a set of tweets collected by the Manchester Guardian newspaper using keyword searches related to the August 6-10, 2011 UK riots and subsequently selected for alluding to false riot-related rumours (tigers freed from London Zoo, London Eye on fire, the army at Bank, police beating teenage girl started riots, Birmingham Children's Hospital attacked, Miss Selfridges on fire, Tottenham rioters cooking food in McDonald's) coded by 2-3 independent coders at the UK National Centre for Text Mining using a single negative to positive polarity scale: -2, -1, 0, 1, 2. The tweets were automatically filtered to remove exact duplicates and manually filtered to remove foreign tweets and near duplicates, such as tweets that were identical except for a user name or the presence or absence of a single hashtag character. Near duplicate removal was important because of the need to split the corpus into training and testing sets. This resulted in a total of 1,698 different tweets (1,342 were removed).

The *AV referendum corpus* is a collection of tweets about the May 5, 2011 Alternative Vote (AV) referendum in the UK. This was a UK-wide vote about whether to replace the existing first past the vote system used for UK parliamentary elections with an AV system in which voters would rank candidates by preference rather than picking a single one. The set was collected by the UK National Centre for Text Mining in order to investigate reactions to the AV referendum and coded for sentiment by 2-3 independent coders from the centre using the same scale as above: -2, -1, 0, 1, 2. This was also filtered to remove duplicates and near duplicates, resulting in 17,963 different tweets (1,795 were removed).

[2] For example, see a newspaper article about one corpus analysis: http://www.guardian.co.uk/uk/2011/dec/07/how-tweets-analysed-understand-riots

The scales used in the above coding exercise were converted to the SentiStrength -4 to +4 scale by averaging the coder scores, multiplying by 2, and rounding to the nearest integer (rounding up in the few cases of .5 decimals – the majority of the texts had three coders).

Each corpus was split equally into two at random: a training set and a test set. To test the mood parameter, the training set was evaluated to decide whether the negative mood would be set and then the test set was used with and without the selected mood parameter to see whether the selection was significant and correct. To test the lexicon extensions two different strategies were used. In both cases term selection was based on the training corpus and performance assessment was based on the test corpus.

To test the technical value of adding the selected terms to improve performance, the top 50 words in terms of occurring in sentences with different human and machine coded values (i.e., *n*=50 in the description above) in the training corpus were selected and manually filtered to remove non-sentiment terms.

To test the human term selection aspect, three independent human coders were chosen. The first author was chosen as someone familiar with how the system works and two other people not associated with the project were chosen as more general system users (an undergraduate linguist and an English graduate). These people were given the lists of terms and information about the topic from the training corpus (to avoid the risk of overfitting) and asked to select the terms from the list that they believed would be typically associated with positive or negative sentiment for the topic but were not nouns or other terms associated with important aspects of the topic[3]. The topic-specific terms were then assessed to see whether they individually improved performance on the test corpus and then the performance of SentiStrength was assessed with the original term list, with the human-selected term list, and with the performance improving human-selected term list (see Appendix, tables 5 and 6). The human coders for the term lists were different from the people used to annotate the training and test corpora.

As a secondary test to assess the scope of the new features, they were also evaluated on six corpora previously generated: MySpace comments, Tweets, YouTube comments, BBC Forum posts, Digg comments, RunnersWorld forum comments (Thelwall et al., 2012). Whilst the first three seem to be very general, two have at least a broad topic focus (news for the BBC and Digg) and one has a narrower topic focus (marathon running for RunnersWorld). These corpora use the dual positive (1 to 5), negative (-1 to -5) human coded scales and the methods described above are modified for this. Two of the human coders for the term lists were again different from the people used to annotate the training and test corpora but the third coder had annotated all the corpora a year previously.

Pearson correlations between the human coded value and the SentiStrength prediction for each text were used to assess sentiment strength detection performance. Correlation is better than calculations based upon raw accuracy (precision, recall and F-measure, as normally used in sentiment polarity or subjectivity detection) when scale rather than binary values are to be tested because it takes into account whether an inaccurate prediction is close to the correct value or not (Strapparava & Mihalcea, 2008; Thelwall et al., 2010). For example, compared to a human value of 4 SentiStrength predictions of -4, -3, -2, -1, 0, 1, 2 or 3 are all incorrect and hence count the same in a precision, recall or F-measure

[3] Template with instructions for coders available at: http://cybermetrics.wlv.ac.uk/paperdata/TermSelectionTopicsPaper.zip

calculation, but 3 is clearly the best prediction and -4 the worst. Correlation calculations take into account the closeness of a prediction because they are partly based upon the *difference* between the two values (predicted and human estimates in this case). For the data sets using the dual scale for positive and negative sentiment, separate correlations were calculated for positive and negative sentiment strength. Alternative sentiment strength measures include "distance precision" (Lu, Kong, Quan, Liu, & Xu, 2010) and mean absolute error (e.g., Thelwall et al., 2010) but neither of these have clear advantages over correlation and correlation has the advantage in a social science context that it is a well-known statistical measure.

This paper compares the performance of SentiStrength with and without the two proposed extensions. Previous articles have evaluated SentiStrength against machine learning approaches (Thelwall et al., 2010; Thelwall et al., 2012), finding similar levels of performance, and these tests are not repeated here because the face validity issues discussed above make them not relevant. No attempt is made here to compare the two SentiStrength extensions with the methods discussed above for domain transfer in machine learning because machine learning is not relevant and these methods are not naturally adaptable for the SentiStrength lexical method. For example, an ensemble of SentiStrength classifiers would be mathematically almost equivalent to a single SentiStrength with average sentiment scores, which is essentially the base version of SentiStrength. In contrast, there seems to be no natural way to adapt structural correspondence learning to the SentiStrength approach since comparing document similarity in a sensible way would be based upon feature vectors that would include terms outside of the sentiment lexicon and hence would undermine the face validity of the results, as discussed above.

Results

Tables 1 and 2 report the results of the two new methods applied to the two main corpora and the secondary corpora. Both training and test corpora values are reported for completeness but the important results are only those for the test corpora.

Table 1. Results of the two mood values applied to the training and test corpora. The highest values are in bold.

Corpus	Training corpus size	Test Corpus size	Training correlation *positive* mood	Training correlation *negative* mood	Test correlation *positive* mood	Test correlation *negative* mood
Riots	847	846	0.3603	**0.4348**	0.3243	**0.4104**
AV	8846	8847	**0.4152**	0.3214	**0.4038**	0.3023
MySpace	520	521	0.6285(+) **0.6089(-)**	**0.6301(+)** 0.5043(-)	**0.5919(+)** **0.6023(-)**	0.5886(+) 0.5308(-)
BBC	500	500	0.3197(+) **0.5696(-)**	**0.3226(+)** 0.5633(-)	**0.3357(+)** **0.6098(-)**	0.3184(+) 0.5912(-)
Digg	538	539	0.4378(+) **0.5460(-)**	**0.4430(+)** 0.5400(-)	**0.3566(+)** **0.5709(-)**	0.3239(+) 0.5465(-)
Runners World	523	523	**0.5452(+)** **0.5318(-)**	0.5353(+) 0.5134(-)	**0.6318(+)** **0.5632(-)**	0.6041(+) 0.5301(-)
Twitter	1173	1174	**0.5257(+)** **0.5402(-)**	0.5085(+) 0.4829(-)	**0.6027(+)** **0.5160(-)**	0.5808(+) 0.4650(-)
YouTube	1704	1703	**0.6100(+)** **0.5391(-)**	0.6012(+) 0.4936(-)	**0.5932(+)** **0.5498(-)**	0.5769(+) 0.4964(-)

The new mood method gives substantial improvement for riots corpus – a 27% increase in correlation from 0.3243 to 0.4104 for the test set. The method gives no change for others since none have overall increase in correlation for negative mood on the training set (after totalling the negative sentiment difference with the positive sentiment difference) and so none would adopt the negative mood on the test set. To check whether this result is sensitive to the random splits used, all the tests in Table 1 were repeated 100 times with different random splits. In all cases except one (1 of the 100 tests for the BBC corpus) the outcome was the same: the test corpora indicating that the negative mood is best overall for the riots corpus and the positive mood is best overall for the other corpora. Tables 2 to 4 give results for the optimal mood for each corpus so that the results of the two methods are reported cumulatively. This only affects the riots corpus, which is reported with negative mood in tables 2 to 4.

Table 2. Results of adding the topic-specific sentiment terms to the training and test corpora (first coder). The highest values for each corpus are in bold.

Corpus	Training correlation original	Training correlation all terms	Training correlation improving terms	Test correlation original	Test correlation all terms	Test correlation improving terms
Riots	0.4348	0.3515	**0.4475**	0.4104	**0.4429**	0.4383
AV	0.4152	0.4124	**0.4255**	0.4038	0.4124	**0.4126**
MySpace	0.6285(+) 0.6089(-)	0.6134(+) 0.5963(-)	**0.6338** **0.6463**(-)	0.5919(+) 0.6023(-)	**0.6134(+)** 0.5963(-)	0.6091(+) **0.6041(-)**
BBC	0.3197(+) 0.5696(-)	0.3376(+) **0.6095**(-)	**0.3427**(+) 0.5857(-)	0.3357(+) 0.6098(-)	**0.3376(+)** 0.6095(-)	**0.3376(+)** **0.6104(-)**
Digg	0.4378(+) 0.5460(-)	0.3554(+) 0.5715(-)	**0.4615**(+) **0.5840**(-)	**0.3566(+)** 0.5709(-)	0.3554(+) **0.5715**(-)	0.3554(+) 0.5709(-)
Runners World	0.5452(+) 0.5318(-)	**0.6305**(+) **0.5632**(-)	0.5616(+) 0.5520(-)	**0.6318(+)** **0.5632(-)**	0.6305(+) **0.5632(-)**	**0.6318(+)** **0.5632(-)**
Twitter	0.5257(+) 0.5402(-)	**0.6024**(+) 0.5160(-)	0.5361(+) **0.5490**(-)	**0.6027(+)** **0.5160(-)**	0.6024(+) **0.5160(-)**	0.6024(+) **0.5160(-)**
YouTube	0.6100(+) 0.5391(-)	0.5878(+) 0.5461(-)	**0.6238**(+) **0.5612**(-)	**0.5933(+)** **0.5498(-)**	0.5878(+) 0.5461(-)	0.5878(+) 0.5462(-)

The human selected values from the first coder (the first author) gave a reasonable improvement for the riots corpus (8% or 7%, depending on the method) and some improvement for the MySpace (3% overall for both methods), alternative votes (2% for both methods) and BBC (1% for both methods) corpora, a decrease in accuracy for the YouTube corpus (-1% for both methods) and no real change for the other four corpora (average of 0% to 0DP). The results for the second coder (Table 3) were similar, with the main exception that the riots corpus with all terms added showed a decrease rather than increase in performance. The results for the third coder (Table 4) were also similar, but showed substantial increases for both the riots and AV corpora for the improving terms, which were the best performing overall for these corpora. The third coder showed a significant decrease in performance for the Runners World corpus[4].

[4] List of terms and those selected by the coders available at: http://cybermetrics.wlv.ac.uk/paperdata/TermSelectionTopicsPaper.zip

Table 3. Results of adding the topic-specific sentiment terms to the training and test corpora (results from the second coder). The highest values for each corpus are in bold.

Corpus	Training correlation original	Training correlation all terms	Training correlation improving terms	Test correlation original	Test correlation all terms	Test correlation improving terms
Riots	0.4348	0.3357	**0.4530**	0.4104	0.3626	**0.4179**
AV	0.4152	0.4216	**0.4223**	0.4038	**0.4082**	**0.4082**
MySpace	0.6285(+) 0.6089(-)	**0.6310**(+) 0.6105(-)	**0.6310**(+) **0.6199**(-)	0.5919(+) 0.6023(-)	0.6176(+) 0.6012(-)	**0.6183**(+) **0.6025**(-)
BBC	0.3197(+) 0.5696(-)	**0.3246**(+) **0.5852**(-)	**0.3246**(+) 0.5824(-)	**0.3357(+)** **0.6098(-)**	**0.3357**(+) 0.6070(-)	**0.3357**(+) 0.6091(-)
Digg	0.4378(+) 0.5460(-)	**0.4622**(+) 0.6024(-)	0.4489(+) **0.5794**(-)	0.3566(+) 0.5709(-)	**0.3577**(+) **0.5752**(-)	**0.3577**(+) 0.5741(-)
Runners World	0.5452(+) 0.5318(-)	**0.5551**(+) **0.5524**(-)	**0.5551**(+) 0.5524(-)	**0.6318**(+) **0.5632**(-)	**0.6318**(+) **0.5632**(-)	**0.6318**(+) **0.5632**(-)
Twitter	0.5257(+) 0.5402(-)	**0.5329**(+) **0.5506**(-)	0.5315(+) 0.5490(-)	0.6027(+) **0.5160**(-)	**0.6106**(+) 0.5153(-)	**0.6106**(+) **0.5160**(-)
YouTube	0.6100(+) 0.5391(-)	0.6198(+) **0.5639**(-)	**0.6227**(+) **0.5639**(-)	**0.5933**(+) 0.5498(-)	0.5856(+) 0.5497(-)	0.5879(+) **0.5501**(-)

Table 4. Results of adding the topic-specific sentiment terms to the training and test corpora (results from the third coder). The highest values for each corpus are in bold.

Corpus	Training correlation original	Training correlation all terms	Training correlation improving terms	Test correlation original	Test correlation all terms	Test correlation improving terms
Riots	0.4348	0.3724	**0.4928**	0.4104	0.3977	**0.4573**
AV	0.4152	0.4126	**0.4455**	0.4038	0.3957	**0.4260**
MySpace	0.6285(+) 0.6089(-)	0.6225(+) 0.6504(-)	**0.6422**(+) **0.6592**	0.5919(+) **0.6023(-)**	**0.6101**(+) 0.5965	**0.6101**(+) 0.5977
BBC	0.3197(+) 0.5696(-)	0.3516(+) **0.6020**(-)	**0.3597**(+) **0.6020**(-)	0.3357(+) **0.6098(-)**	0.3385(+) **0.6098**(-)	**0.3403**(+) **0.6098**(-)
Digg	0.4378(+) 0.5460(-)	0.4883(+) 0.6085(-)	**0.4992**(+) **0.6123**(-)	**0.3566(+)** 0.5709(-)	0.3538(+) 0.5688(-)	0.3549(+) **0.5712**(-)
Runners World	**0.5452(+)** 0.5318(-)	0.5312(+) **0.5529**(-)	0.5312(+) **0.5529**(-)	**0.6318(+)** **0.5632(-)**	0.6119(+) 0.5616(-)	0.6119(+) 0.5616(-)
Twitter	0.5257(+) 0.5402(-)	0.5299(+) 0.5496(-)	**0.5378**(+) **0.5534**(-)	0.6027(+) 0.5160(-)	**0.6061**(+) **0.5193**(-)	0.6027(+) 0.5191(-)
YouTube	0.6100(+) 0.5391(-)	0.6231(+) 0.5709(-)	**0.6240**(+) **0.5743**(-)	**0.5933(+)** 0.5498(-)	0.5880(+) **0.5513**(-)	0.5880(+) 0.5498(-)

Limitations

Probably the most important limitation for the two main corpora is that Twitter seems to encourage duplication of content, in the form of retweets, and that despite the duplicate and near duplicate filtering the main two corpora used will include some tweets in the

training half that are similar to tweets in the test half. This makes the training/testing dichotomy imperfect. This seems unlikely to impact the mood results but may impact the lexicon extension results. A second issue is that the riot rumours corpus was collected in an unknown way by an external organisation and so it is not certain that it is not biased in some way, presumably by using a set of keyword queries. These issues should not impact the six secondary corpora, however, since these are all small samples taken from very large collections and include date collected using methods under the control of the researcher.

An additional limitation is that the issue addressed in this article is by its nature topic-specific and it seems likely that there will be topics for which the new methods will not work well. Hence the results should not be taken as evidence that the methods will always work for specific topics, but that they work for *some* topics.

Discussion and conclusions

Overall, the mood method made a significant improvement in the results for the riots corpus but did not change the results for the other corpora. A switch to a negative overall mood gave a significant decrease in performance for the alternative vote corpus and so it is recommended not to always use one mood but to choose the mood based upon the corpus. The results thus suggest that identifying mood based upon human classification of a training set of data will in some cases improve sentiment analysis performance. If a training set is not available then it seems intuitively likely that a negative mood should only be used for a clearly negative topic, but many more topics would need to be analysed to fully assess this rule.

The lexical extension method gave a smaller improvement than the mood method but an increase of 8% in the case of one corpus nevertheless represents a substantial improvement in sentiment analysis terms. Hence this method is recommended even though it may occasionally slightly worsen performance. There was little to choose between the method of adding all human-filtered terms and the method of just adding the human filtered terms that produce an improvement on the test corpus but the latter is nevertheless recommended to safeguard against accidentally adding a poor term.

The lexical extension method gave two unexpected results. The riots corpus exhibited a substantial decrease in performance on the training corpus when the human selected terms are added. This appears to be due the single word *fire* which occurred in several posts in the training corpus with the apparently joke theme of pointing to the URL of a picture of the person who "set fire to Miss Selfridges" – some of these were coded as neutral unless the person in question was also insulted in the post (e.g., scumbag). Hence, the word fire was used in a context that the human term coders may not have expected.

The second unexpected result was that the lexical extension method did not improve performance on the RunnersWorld posts, despite them having a fairly narrow topic. After excluding typos or deliberately unusual spellings, the list of improving terms for this forum was small: sweaty (-2), whacked (-2), crapping (-2), magic (2), hydrate (2), achievable (2), and fair (2). Hence, despite its narrow focus, it seems that the forum does not use a large specialist vocabulary but mainly uses terms that are already in the SentiStrength lexicon.

The human filtering step in the lexical method is important, even though it may reduce overall classification performance. For example, in the AV corpus, four of the top terms were *no*, *no2av*, *yes* and *yes2av*. If these were added to the lexicon with sentiment strengths of -2, -2, 2, and 2 respectively then the correlations would increase to 0.4726 (training corpus) and 0.4637 (test corpus). These significant improvements would come at

the expense of a loss in validity for researchers concerned to contrast sentiment expressions between the two campaigns, however.

In summary, both the mood and lexical extension methods have been shown to improve the performance of sentiment strength detection in some case and are therefore recommended for cases where social web text with a particular topic focus is analysed. Whilst both methods can be applied to any context, the lexical extension method is particularly suited to social web data that is focused on a particular topic. Both methods require human intervention – the first to annotate a corpus and an additional small amount for the second to also select terms. The use of human labour seems unavoidable in the context of the need for high face validity for social science applications, however, despite the emergence of more automated methods that are increasingly able to differentiate between sentiment and non-sentiment contexts (e.g., Zhang, & Liu, 2011). Human labour seems likely to be particularly important for narrowly-focused topics for which small misclassifications may result in significant discrepancies if they are for terms that are frequently used with regard to a key aspect of the topic.

Despite the improvements shown for one corpus, one data type analysed, BBC forum discussions, still gives mediocre results for positive sentiment strength detection compared to the other corpora and needs additional sentiment analysis methods to be developed for it. For example modifications of aspect-based sentiment analysis methods may improve performance for specific topics. It would also be useful to test the mood methods on a range of negative topics to assess whether moods should always be negative for negative topics. Finally, it would also be interesting to try the lexical extension method for larger training corpora and presumably this would show that it is more effective with more training data.

Acknowledgement

This work was supported by a European Union grant by the 7th Framework Programme, Theme 3: Science of complex systems for socially intelligent ICT. It is part of the CyberEmotions project (contract 231323). It was also supported by JISC as part of the *Twitter analysis workbench development* project. Thank you to Rob Procter and the National Centre for Text Mining in Manchester for supplying and coding the Riots and AV corpora.

Appendix

Table 5. Topic-specific sentiment terms for the riots corpus from the first coder. Selected terms made an individual improvement on the training corpus.

Term	Weight	Selected
arrest	-2	
arrested	-2	
baton	-2	x
batoned	-3	x
birminghamriots	-2	
brainwashing	-3	x
caught	-2	
demanded	-2	
fire	-3	
gutted	-3	x
helpharryhelpothers	2	
hospital	-2	
londonriots	-2	x
londonriots2011	-2	
londonzoobreakin	-2	
manchesterriots	-2	
pigs	-3	x
pls	2	
protect	2	x
rioters	-3	
roaming	-2	x
rooters	-2	
rumors	-2	
rumour	-2	
suspected	-2	

Table 6. Topic-specific sentiment terms for the alternative vote corpus from the first coder. Selected terms made an individual improvement on the training corpus.

Term	Weight	Selected
ace	3	
ass	-2	
better	2	x
cut	-2	x
fairer	2	x
fearmongerers	-3	
feck	-3	
mongering	-3	
moral	2	x
naysayers	-2	
scruffy	-2	x
viva	2	x

References

Agerri, R., & García-Serrano, A. (2010). Q-WordNet: Extracting polarity from WordNet senses. *Proceedings of the Seventh conference on International Language Resources and Evaluation*, Retrieved May 25, 2010 from: http://www.lrec-conf.org/proceedings/lrec2010/pdf/2695_Paper.pdf.

Aue, A., & Gamon, M. (2005). Customizing sentiment classifiers to new domains: A case study. *Proceedings of the International Conference on Recent Advances in Natural Language Processing*, Retrieved March 6, 2011 from: http://research.microsoft.com/pubs/65430/new_domain_sentiment.pdf.

Baccianella, S., Esuli, A., & Sebastiani, F. (2010). SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. *Proceedings of the Seventh conference on International Language Resources and Evaluation*, Retrieved May 25, 2010 from: http://www.lrec-conf.org/proceedings/lrec2010/pdf/2769_Paper.pdf.

Blitzer, J., McDonald, R., & Pereira, F. (2006). Domain adaptation with structural correspondence learning. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP 2006)* (pp. 120-128).

Chung, J. E., & Mustafaraj, E. (2011). Can collective sentiment expressed on Twitter predict political elections? In W. Burgard & D. Roth (Eds.), *Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence (AAAI 2011)* (pp. 1768-1769). Menlo Park, CA: AAAI Press.

Ding, X., Liu, B. & Yu, P.S. (2008). A holistic lexicon-based approach to opinion mining. In: Proceedings of the Conference on Web Search and Web Data Mining (WSDM-2008) New York: ACM Press (pp. 231-240).

Dodds, P. S., & Danforth, C. M. (2010). Measuring the happiness of large-scale written expression: Songs, blogs, and presidents. *Journal of Happiness Studies, 11*(4), 441-456.

Glorot, X., Bordes, A., & Bengio, Y. (2011). Domain adaptation for large-scale sentiment classification: A deep learning approach. In *Proceedings of the 28 th International Conference on Machine Learning (ICML 2011)*.

Gruzd, A., Doiron, S., & Mai, P. (2011). Is happiness contagious online? A case of Twitter and the 2010 Winter Olympics. In *Proceedings of the 44th Hawaii International Conference on System Sciences* (pp. Retrieved June 2, 2011 from: http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?tp=&arnumber=5718715). Washington, DC: IEEE Computer Society.

Guardian. (2012). Data journalism and data visualisation. http://www.guardian.co.uk/data.

Hatzivassiloglou, V. & McKeown, K.R. (1997). Predicting the semantic orientation of adjectives. In: P. Cohen & W. Wahlster (eds). Proceedings of Annual Meeting of the Association for Computational Linguistics (ACL-1997). Madrid, Spain 7-12 July 1997 (pp. 174–181).

Hu, M., & Liu, B. (2004). Mining and summarizing customer reviews. *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '04)*, ACM Press: New York, NY (p. 168-177).

Kaji, N. & Kitsuregawa, M. (2007). Building lexicon for sentiment analysis from massive collection of HTML documents. In: Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL 2007) June 28-30, Prague, Czech Republic (pp. 1075-1083).

Kramer, A. D. I. (2010). An unobtrusive behavioral model of "Gross National Happiness". In *Proceedings of CHI 2010* (pp. 287-290). New York: ACM Press.

Liu, B. (2012). *Sentiment analysis and opinion mining*. San Rafael, CA: Morgan and Claypool.

Lu, Y., Kong, X., Quan, X., Liu, W., & Xu, Y. (2010). Exploring the sentiment strength of user reviews. *Lecture Notes in Computer Science, 6184/2010*, 471-482.

Lu, Y., Castellanos, M., Dayal, U., & Zhai, C. (2011). Automatic construction of a context-aware sentiment lexicon: An optimization approach, In: Proceedings of the 20th international conference on World wide web (WWW'2011) New York: ACM Press (pp. 347—356).

Mejova, Y. & Srinivasan, P. (2011). Exploring feature definition and selection for sentiment classifiers. In: *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media* (ICWSM-2011), Menlo Park, CA: AIII Press (pp. 546-549).

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

Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up? Sentiment classification using machine learning techniques. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (pp. 79-86). Morristown, NJ: Association for Computational Linguistics.

Pennebaker, J., Mehl, M., & Niederhoffer, K. (2003). Psychological aspects of natural language use: Our words, our selves. *Annual Review of Psychology, 54*, 547-577.

Ponomareva, N., & Thelwall, M. (2012). Do neighbours help? An exploration of graph-based algorithms for cross-domain sentiment classification. *The 2012 Conference on Empirical Methods on Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL 2012)* (pp. 655-665).

Riloff, E., Patwardhan, S., & Wiebe, J. (2006). Feature subsumption for opinion analysis. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 440-448.

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

Strapparava, C., Valitutti, A., & Stock, O. (2006). The affective weight of lexicon. *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006)*, Retrieved July 28, 2011 from: http://gandalf.aksis.uib.no/lrec2006/pdf/2186_pdf.pdf.

Strapparava, C., & Mihalcea, R. (2008). Learning to identify emotions in text, *Proceedings of the 2008 ACM symposium on Applied computing* (pp. 1556-1560). New York, NY: ACM.

Taboada, M., Anthony, C., & Voll, K. (2006). Methods for creating semantic orientation dictionaries. In *Proceedings of Fifth International Conference on Language Resources and Evaluation (LREC 2006)* (pp. 427-432). Genoa, Italy: ELRA.

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational Linguistics, 37*(2), 267-307.

Tan, S., Wu, G., Tang, H., & Cheng, X. (2007). A novel scheme for domain-transfer problem in the context of sentiment analysis. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management (CIKM 2007)* (pp. 979-982). New York, NY: ACM Press.

Thelwall, M., Buckley, K., & Paltoglou, G. (2012). Sentiment strength detection for the social Web. *Journal of the American Society for Information Science and Technology, 63*(1), 163-173.

Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., & Kappas, A. (2010). Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology, 61*(12), 2544-2558.

Turney, P. D. (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 (ACL), July 6-12, 2002, Philadelphia, PA*, 417-424.

Wilson, T. (2008). *Fine-grained subjectivity and sentiment analysis: Recognizing the intensity, polarity, and attitudes of private states.* University of Pittsburgh.

Wu, Q., Tan, S., & Cheng, X. (2009). Graph ranking for sentiment transfer. In *Proceedings of the ACL-IJCNLP 2009 Conference (ACL-IJCNLP '09)* (pp. 317–320).

Zagibalov, T. (2010). *Unsupervised and knowledge-poor approaches to sentiment analysis.* University of Sussex, Brighton.

Zhang, L. & Liu, B. (2011). Identifying noun product features that imply opinions. In Proceedings of the Annual Meeting of the Association for Computational Linguistics – Human Language Technologies (HLT '11), ACL Press: Stroudsburg, PA, (pp. 575-580).