Sentiment Strength Detection for the Social Web[1]

Mike Thelwall, Kevan Buckley, Georgios Paltoglou
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, G.Paltoglou@wlv.ac.uk
Tel: +44 1902 321470 Fax: +44 1902 321478

Sentiment analysis is concerned with the automatic extraction of sentiment-related information from text. Although most sentiment analysis addresses commercial tasks, such as extracting opinions from product reviews, there is increasing interest in the affective dimension of the social web, and Twitter in particular. Most sentiment analysis algorithms are not ideally suited for this task because they exploit indirect indicators of sentiment that can reflect genre or topic instead. Hence, such algorithms used to process social web texts can identify spurious sentiment patterns caused by topics rather than affective phenomena. This article assesses an improved version of the algorithm SentiStrength for sentiment strength detection across the social web that primarily uses direct indications of sentiment. The results from six diverse social web data sets (MySpace, Twitter, YouTube, Digg, Runners World, BBC Forums) indicate that SentiStrength 2 is successful in the sense of performing better than a baseline approach for all data sets in both supervised and unsupervised cases. SentiStrength is not always better than machine learning approaches that exploit indirect indicators of sentiment, however, and is particularly weaker for positive sentiment in news-related discussions. Overall, the results suggest that, even unsupervised, SentiStrength is robust enough to be applied to a wide variety of different social web contexts.

Introduction

Whilst sentiment analysis often focuses on reviews of movies or consumer products (Gamon, Aue, Corston-Oliver, & Ringger, 2005; Tang, Tan, & Cheng, 2009), these probably form a tiny fraction of the social web. The remainder includes many friendly exchanges in social network sites (SNSs), discussions of politics, sport and the news in blogs and online forums as well as comments on media published in YouTube, Flickr and Last.FM. Analysing sentiment in this much broader class of text is valuable from a social sciences perspective because it can aid the discovery of sentiment-related patterns, such as gender differences and successful communication strategies. For instance such analyses have shown that females give and receive stronger positive sentiments than males in the SNS MySpace (Thelwall, Wilkinson, & Uppal, 2010), the existence of sentiment homophily in SNSs (Bollen, Gonçalves, Ruan, & Mao, 2011; Thelwall, 2010), that sentiment is important in online groups formed around blogs (Mitrovic, Paltoglou, & Tadic, 2011) and that initial negative sentiments help to generate longer online discussions (Chmiel et al., 2011; Naveed, Gottron, Kunegis, & Alhadi, 2011). However, analysing social web texts using traditional sentiment analysis methods for social science research is problematic for several reasons.

The first issue is that human coded data (i.e., a set of texts assessed by humans for sentiment) must be manually created since non-review texts are rarely annotated for sentiment by the author or readers (for exceptions see: Mishne, 2005; Mishne & de Rijke, 2006). This human coded data is needed to assess the accuracy of all sentiment analysis algorithms and as an input to train most machine learning sentiment analysis algorithms. Second, sentiment analysis is known to be domain-dependent, meaning that applying a classifier to a data set different from the one on which it was trained often gives poor results (Aue & Gamon, 2005). The diversity of topics and communication styles in the social web suggests that many different classifiers may be needed.

Most seriously for some purposes, classifiers that are technically optimised to a domain (i.e., having the highest accuracy scores) may use *indirect* indicators of sentiment and therefore give misleading results for social science research by identifying spurious patterns. For instance, a trained classifier for political discussions is likely to learn words like Iraq, Iran, Palestine and Israel as strong indicators of negativity since, in a political context, these are typically associated with bad news, strong opinions, and heated debates. Such a classifier is implicitly using these terms as (effective) indicators of sentiment, but this obscures the identification of direct expressions of sentiment and can

[1] This is a preprint of an article published in the Journal of the American Society for Information Science and Technology © copyright 2011 John Wiley & Sons, Inc. 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.

be unhelpful for social science attempts to identify patterns of sentiment. Thus an investigation into emotions triggered in news discussions might discover that they most frequently occur in discussions of the Middle East (probably correct but unsurprising) rather than that they tend to occur most strongly at the start of discussions about most topics (a more unexpected and more useful finding). Another clear example of the problem is for research into sentiment homophily (e.g., Bollen, Gonçalves et al., 2011; Thelwall, 2010): studies of the tendency for communication partners or Friends to use similar types of sentiment should not use a machine learning approach since this might indicate that they have topics in common rather than sentiment in common. Other examples include studies of trends in sentiment over time (Diakopoulos & Shamma, 2010; Kramer, 2010; O'Connor, Balasubramanyan, Routledge, & Smith, 2010; Thelwall, Buckley, & Paltoglou, 2011) or emotion contagion (Gruzd, Doiron, & Mai, 2011), which could potentially track trends in topics, such as Palestine, over time. This could particularly impact on sentiment analysis for news (Balahur et al., 2010) or politics (Balahur, Kozareva, & Montoyo, 2009). Some commercial applications of sentiment analysis may also suffer from similar problems (Taboada, Brooke, Tofiloski, Voll, & Stede, 2011), as described below.

From the above, it is sometimes critical to have classifiers that are only allowed to exploit direct indicators of sentiment. This is possible with a lexical approach: i.e., performing the sentiment analysis primarily by identifying the presence of terms from a lexicon of known sentiment-bearing words or phrases. Lexical approaches have been used in many types of sentiment analysis. They typically incorporate sentiment word lists from resources such as the General Inquirer (GI) lexicon (Stone, Dunphy, Smith, & Ogilvie, 1966), the ANEW words (Bradley & Lang, 1999), SentiWordNet (Baccianella, Esuli, & Sebastiani, 2010) or the LIWC dictionary (Pennebaker, Mehl, & Niederhoffer, 2003). Methods have also been developed to automatically create sentiment coded lexicons, such as from the adjectives extracted from a set of texts (Hatzivassiloglou & McKeown, 1997; Taboada et al., 2011; Taboada & Grieve, 2004).

Sentiment can be assessed for polarity – whether it is positive or negative – but can also be assessed for the strength with which a positive or negative sentiment is expressed. The sentiment strength detection task addressed in the current paper involves assessing both the strength of positive sentiment *and* the strength of negative sentiment in a text, with the assumption that both positive and negative sentiment can coexist within texts. Hence a text is given two scores: a positive sentiment strength score and a negative sentiment strength score. An alternative approach is to have a single scale combining sentiment polarity and strength (Taboada et al., 2011). In order to tackle either sentiment strength detection task, sentiment terms can also be associated with default strengths, for example giving *love* a stronger weighting than *like*. This has been used in SentiStrength, which is the focus of the current paper. SentiStrength is designed to identify positive and negative sentiment strength in short informal social web text and has been applied to comments in the SNS MySpace (Thelwall, Buckley, Paltoglou, Cai, & Kappas, 2010). The same approach is used in SO-CAL (Taboada, Brooke, & Stede, 2009), which classifies texts on a single negative to positive scale (Taboada et al., 2011).

This article assesses whether a lexical algorithm that primarily relies upon direct indicators of sentiment, an improved version of the MySpace sentiment strength detection program SentiStrength, is generally effective for positive and negative sentiment strength detection across the social web. This is achieved by testing SentiStrength on human-coded texts from six different social web domains: not only MySpace but also Twitter, YouTube, the Runners World marathon discussion forum, the Digg news identification site and the BBC Forum news discussion site.

Sentiment analysis

The two most common sentiment analysis tasks are subjectivity and polarity detection. The former predicts whether a given text is subjective or not and the latter predicts whether a subjective text is positive or negative overall. Less common is sentiment strength detection, which predicts the strength of positive or negative sentiment within a text. This section primarily deals with polarity detection although the methods are applicable to all three tasks.

A common approach for sentiment analysis is to select a machine learning algorithm and a method of extracting *features* from texts and then train the classifier with a human-coded corpus. The features used are typically words but can also be stemmed words or part-of-speech tagged words, and also may be combined into bigrams (e.g., two consecutive words) and trigrams (Pang & Lee, 2008).

More sophisticated variations have also been developed, such as for intelligent feature selection (Riloff, Patwardhan, & Wiebe, 2006).

An alternative polarity detection method is to identify the likely average polarity of words within texts by estimating how often they co-occur with a set of seed words of known and unambiguous sentiment (e.g., good, terrible), typically using web search engines to estimate relative co-occurrence frequencies (Turney, 2002). The assumption here is that positive words will tend to co-occur with other positive words more than with negative words, and vice-versa. This approach needs relatively little lexical input knowledge and is flexible for different domains in the sense that a small set of initial general keywords can be used to generate a different lexicon for each application domain. The seed words method seems to perform reasonably well in a variety of different contexts and learns domain-specific sentiment-associated words, such as 3G for mobile phones (Zagibalov, 2010).

The fact that machine learning methods are normally domain-specific (i.e., do not work well on topics or text genres that are different from those that they were trained on) has led to interest in domain transfer: methods for generating an effective classifier for a new domain based upon a classifier trained for an old domain, typically using similarities between the new and old domains (Andreevskaia & Bergler, 2008; Tan, Wu, Tang, & Cheng, 2007). For instance, one approach is to use a classifier trained on one domain to identify documents in another domain that can be classified with a high degree of confidence, and then use structure in the new domain to help predict the remaining classifications (Wu, Tan, Duan, & Cheng, 2010). Domain transfer methods reduce the need for human coded data in new target domains but seem to give inferior results to direct training on target domains with sufficient training data (Zagibalov, 2010).

As previously stated, all methods discussed here are likely to identify terms that associate with sentiment but do not directly express it, such as *feel*, *Iraq*, and *late*. Such terms have been called *indirect affective words* to distinguish them from *direct affective words* (Strapparava, Valitutti, & Stock, 2006). The use of indirect affective words is a drawback for some types of social science sentiment analysis research and also for some commercial applications because it makes the methods domain-dependant and sometimes also time dependant (e.g., 3G is probably no longer a reliable indicator of a positive mobile phone reviews).

Lexical algorithms

The lexical approach is to start with an existing set of terms with known sentiment orientation and then use an algorithm to predict the sentiment of a text based upon the occurrences of these words. The lexicon method can be supplemented with other information, such as emoticon lists, and semantic rules, such as for dealing with negation (Neviarouskaya, Prendinger, & Ishizuka, 2007; Taboada et al., 2011). As mentioned above, the lexicon used can be derived from a variety of sources, such as the General Inquirer lexicon (Stone et al., 1966), the ANEW words (Bradley & Lang, 1999), SentiWordNet (Baccianella et al., 2010) WordNet Affect (Strapparava & Valitutti, 2004) or the LIWC dictionary (Pennebaker et al., 2003). Moreover, various methods have been developed to improve on standard sources, such as by detecting compound words (Neviarouskaya, Prendinger, & Ishizuka, 2011). As discussed above for the seed words algorithm, additional terms can improve performance in specific domains and some lexical algorithms are able to learn non-sentiment terms that associate with sentiment in particular domains, such as “small” being a general positive word for portable electronic device reviews (Yue Lu, Castellanos, Dayal, & Zhai, 2011; Velikovich, Blair-Goldensohn, Hannan, & McDonald, 2010).

Although designed for a different task, the program that is internally most similar to that of the current paper, SentiStrength, is SO-CAL, which uses a lexical strategy to code texts as positive or negative. It uses lexicons of terms coded on a single negative to positive scale of -5 to +5 (Taboada et al., 2011). SO-CAL’s lexicon was built by human coders tagging all the adjectives, nouns, verbs and adverbs for strength and polarity in 500 texts from several corpora, as well as the General Inquirer lexicon (Stone et al., 1966). This generated 2,252 adjectives, 745 adverbs 1,142 nouns and 903 verbs, and all nouns and verbs were lemmatised, making the effective list size larger (Taboada et al., 2011). Words were coded for their “prior polarity” – their assumed normal polarity across all contexts - rather than their polarity in the particular context in which they were found. SO-CAL also has at least 187 multi-word sentiment expressions. It has a set of intensifying expressions that increase or decrease the sentiment strength of subsequent words (e.g., *extraordinarily*) and procedures for dealing with negation (motivated by: Polanyi & Zaenen, 2006). Words, such as *would*, that effectively neutralise any sentiment following are also used. SO-CAL boosts the strength of negative expressions

in texts since they seem to be less common than positive expressions, and reduces the strength of terms that occur frequently. The final polarity decision is determined by the average sentiment strengths of the words detected, after modifications. Tests on multiple data sets showed SO-CAL to perform consistently well for polarity detection across a range of balanced data sets with mainly web or news content (Taboada et al., 2011). A program with a similar broad overall approach has also been tested by two of the authors of the current paper on three of the data sets used in the current paper (Paltoglou & Thelwall, in press), with good unsupervised results for both polarity and subjectivity detection in comparison to machine learning.

If the goal is sentiment *strength* detection rather than polarity or subjectivity detection then the lexicon is likely to incorporate human-estimated sentiment weights (Yao Lu, Kong, Quan, Liu, & Xu, 2010; Neviarouskaya et al., 2007). For instance, *ache* might be scored -2 as mildly negative but *excruciating* scored -5 as strongly negative. These scores would help an algorithm to distinguish between weak and strong sentiment in sentences containing these words.

Polarity detection can be conceived as identifying groups of sentiments. For example, positive texts may include expressions of happiness, love, contentment and euphoria, which have different strengths and types. A deeper parsing linguistic sentiment analysis method is to attempt to identify grammatical structure units within sentences and to use this for phrase level sentiment analysis (Wilson, Wiebe, & Hoffman, 2009), fine-grained sentiment classifications (e.g., anger, love, fear) (Neviarouskaya, Prendinger, & Ishizuka, 2010) and opinion intensity (strength) classifications (Wilson, Wiebe, & Hwa, 2006). This may not work well in text that disobeys standard rules of grammar, however, and hence may not work well in parts of the social web in which high levels of informality are common.

Sentiment strength algorithms have also been defined for multiple emotions, using linguistic structure, as described above (Neviarouskaya et al., 2010; Wilson et al., 2006). One study compared a variety of different approaches for sentiment strength detection of news headlines, finding that Naïve Bayes machine learning did not work as well as the use of linguistic information from WordNet, WordNet Affect and SentiWordNet (Strapparava & Mihalcea, 2008). The Naïve Bayes method is not necessarily the best one for this task, however, so this does not prove that machine learning is necessarily inferior to the lexical approach for sentiment strength detection.

Some sentiment analysis algorithms have included special adaptations for the social web. One obvious feature is the use of emoticons to directly express sentiment (Mishne & de Rijke, 2006; Neviarouskaya et al., 2007). Emotions have also been used as sentiment markers to annotate a corpus for machine learning (Pak & Paroubek, 2010; Read, 2005). Other features used include repeated punctuation, words written in all capital letters and standard abbreviations (Neviarouskaya et al., 2007). SentiStrength has also introduced new capabilities, such as the use of repeated letters within a word for sentiment emphasis (Thelwall, Buckley et al., 2010). The complete set of SentiStrength rules is described in the next section.

SentiStrength 2

SentiStrength is a lexicon-based classifier that uses additional (non-lexical) linguistic information and rules to detect sentiment strength in short informal English text. For each text, the SentiStrength output (for both version 1 and version 2) is two integers: 1 to 5 for positive sentiment strength and a separate score of 1 to 5 for negative sentiment strength. Here, 1 signifies no sentiment and 5 signifies strong sentiment of each type. For instance, a text with a score of 3, 5 would contain moderate positive sentiment and strong negative sentiment. A neutral text would be coded as 1, 1. Two scales are used because even short texts can contain both positivity and negativity and the goal is to detect the sentiment expressed rather than its overall polarity (Thelwall, Buckley et al., 2010). Below is a list of SentiStrength's key features (Thelwall, Buckley et al., 2010). Those marked with ^ have been superseded in version 2.

- A **sentiment word list with human polarity and strength judgements^.** Some words include Kleene star stemming (e.g., ador*).
    - The word "**miss**" is a special case with a positive and negative strength of 2. It is frequently used to express sadness and loves simultaneously.
- A **spelling correction algorithm** deletes repeated letters in a word when the letters are more frequently repeated than normal for English or, if a word is not found in an English dictionary, when deleting repeated letters creates a dictionary word (e.g., hellp -> help).

- A **booster word list** is used to strengthen or weaken the emotion of following sentiment words.
- An **idiom list^** is used to identify the sentiment of a few common phrases. This overrides individual sentiment word strengths.
- A **negating word list^** is used to invert following emotion words (skipping any intervening booster words).
- **At least two repeated letters** added to words give a strength boost sentiment words by 1. For instance haaaappy is more positive than happy. Neutral words are given a positive sentiment strength of 2 instead.
- An **emoticon list with polarities** is used to identify additional sentiment.
- Sentences with **exclamation marks** have a minimum positive strength of 2, unless negative.
- **Repeated punctuation** with one or more exclamation marks boost the strength of the immediately preceding sentiment word by 1.
- **Negative sentiment is ignored in questions^**.

There are two versions of SentiStrength: supervised and unsupervised (only the supervised version was discussed in the previous paper). The supervised version has the following additional component.

- A **training algorithm that optimises sentiment word strengths and potentially also changes polarity** (i.e., supervised learning). The algorithm checks each term strength to see whether an increase or decrease of 1 would increase classification accuracy on a corpus of human-classified texts (i.e., training data). The algorithm repeats until all words have been checked without making any changes.

The original version of SentiStrength was only tested on the short informal friendship messages of the SNS MySpace and a new version was developed to cope with a wider variety of types of text. The main change is a significant extension of the lexicon for negative terms by the incorporation of the negative General Inquirer terms (Stone et al., 1966). This extension (called SentiStrength 2) was designed to address SentiStrength's relatively weak performance for negative sentiment strength detection. In particular:

- The sentiment word list was extended with negative GI terms with human-coded sentiment weights and Kleene star stemming. This increased the number of terms in the sentiment word list from 693 to 2310.
- The sentiment word terms was tested against a dictionary to check for incorrectly matching words and derivative words that did not match. This resulted in many terms being converted to wildcards (e.g., to match –ness word variants) and some exclusions being added (e.g., amazon* added as an exclusion for amaz*, admiral* added as an exclusion for admir*). Exclusions were typically rare words matching common sentiment words but longer). SentiStrength was recoded to match the longest term if multiple terms matched. This increased the sentiment word list to 2489 terms, 228 of which were neutral (strength 1), either as exclusions or as potential sentiment words that could be incorporated by the training stage. Most (1364) terms had a Kleene star ending after this stage.
- Negating negative terms makes them neutral rather than positive (e.g., "I do not hate him", is not positive).
- The idiom list was extended with phrases indicating word senses for common sentiment words. For instance, "is like" has strength 1 (the minimum score on the positive scale, indicating neutral text) because "like" is a comparator after "is" rather than a positive term (strength 2). This is a simple alternative to part of speech tagging for the most important sentiment word contexts relevant to the algorithm scores.
- The special rule for negative sentiment in questions was removed.

Research questions

The goal of this study is to assess SentiStrength 2 in a variety of different online contexts to see whether it is a viable as a general sentiment strength detection algorithm for the social web, despite its primary reliance upon direct affective terms. Since viability is the goal rather than optimal performance and the task is sentiment strength detection, the requirement is that SentiStrength 2 results should have a statistically significant positive correlation with both positive and negative sentiment on all data sets. Ideally, this should be true for the unsupervised version of SentiStrength 2 that does not need training data because this would mean that the task of creating human coded data

sets for each social web context to train the algorithm would be unnecessary. A secondary goal is to assess how well SentiStrength 2 performs in comparison to other methods that exploit indirect affective terms and which types of social web data it performs best on. Standard machine learning methods are used for this comparison since no other programs than SentiStrength or similar algorithms perform this task. The following questions are therefore addressed.

- Does the unsupervised version of SentiStrength 2 give a significant positive correlation with all types of social web texts for both positive and negative sentiment?
- Does the supervised version of SentiStrength 2 give a significant positive correlation with all types of social web texts for both positive and negative sentiment?
- Does SentiStrength 2 perform better than standard machine learning algorithms on social web texts?

Methods and data

SentiStrength 2 was tested on the following six human-coded data sets, plus a combined data set containing all of them. These were chosen to represent a variety of different types of public social web environment. The list is not exhaustive, however. For example it excludes chat environments and newsgroups.

- BBC Forum posts: Public news-related discussions. This represents discussions about various serious topics, from national and world news to religion and politics.
- Digg.com posts: Public comments on news stories. This represents general news commentary and evaluation.
- MySpace comments: Public messages between Friends in this SNS. This data represents SNS communication.
- Runners World forum posts: Public group messages on the topic of marathon running. This data represents specialist forums for common-interest groups.
- Twitter posts: Public microblog broadcasts. Twitter is an important site in its own right.
- YouTube comments: Text comments posted to videos on the YouTube web site. This represents comments on resources and any associated discussions.
- All six combined: All of the above were combined into a single large data set to assess how well SentiStrength 2 performed in a mixed environment and to see whether a significant increase in training data would give a large relative increase in the performance of the selected machine learning methods.

The texts in each data source were coded over 20 hours (a maximum of 1 hour per day) by 1-3 different people operating independently but using a common code book (see: Thelwall, Wilkinson et al., 2010). The coders were selected from an initial set of 9 people for consistent results and were allowed to use their own judgements rather than being trained to code in a pre-defined way. The data sets were coded by three people (using the average score in each case, rather than discarding texts with disagreement) except for Runners World (2, with a third as an arbitrator for ties), Twitter (1) and YouTube (1). None of the coders were otherwise involved in the research and none were sentiment analysis researchers. Krippendorff's α was used to assess inter-coder reliability because it can cope with multiple coders and ordinal categories (Artstein & Poesio, 2008; Krippendorff, 2004). Numerical differences in sentiment score were used as the weights for this metric. For positive sentiment, the α values were 0.5743 (MySpace), 0.4293 (BBC), 0.5011 (Digg) and 0.6809 (Runners World for the two coders). For negative sentiment, the α values were 0.5634 (MySpace) 0.5068 (BBC), 0.4910 (Digg) and 0.6623 (Runners World for the two coders). These values indicate moderate agreement: the coders had broadly similar but not identical perceptions of sentiment.

A range of standard machine learning algorithms were selected to compare against SentiStrength 2 and each was assessed on a set of different features and feature set sizes as in the previous SentiStrength paper (Thelwall, Buckley et al., 2010). Stopwords were not removed because common words, such as "I" and "you", can associate with expressions of sentiment. The algorithms used were: support vector machines (Sequential Minimal Optimization variant, SMO), Logistic Regression (SLOG for short), ADA Boost, SVM Regression, Decision Table, Naïve Bayes, J48 classification tree, and JRip rule-based classifier. The previously selected Multilayer Perceptron algorithm was not used as it performed poorly and was very slow. The processing was conducted by Weka (Witten & Frank, 2005). The subsumption technique for improving machine learning feature selection (Riloff et al., 2006) was not used as it did not improve performance in previous tests with

social web data. As an additional check, however, SVM regression with subsubmption (alpha = 0.05, 0.1 and 0.2) was applied to all the data sets via the commonly used SVM light (Joachims, 1999) but it was outperformed in all cases. Although the main performance measure of interest is correlation, accuracy (i.e., the number of times that the computer prediction is exactly the human-coded value) and accuracy ± 1 class were also calculated for additional evidence. The correlation used was the Pearson coefficient, calculated on a given text set between the values produced by the algorithm and the human-coded values.

Each algorithm was tested 30 times using 10-fold cross validation with 10 different feature set sizes (100, 200,… 1000) and the best algorithm and feature set size was reported for each data set. More specifically, the algorithm reported was the one with the highest correlation (calculated as above) averaged over the 30 repetitions. This use of a wide variety of algorithms and feature sets tested gives the machine learning approach in general an "unfair" advantage over SentiStrength because some algorithms are statistically likely to perform better than normal due to random factors within the data.

The feature set used for the machine learning was made more powerful than in previous experiments (Thelwall, Buckley et al., 2010) by using the emoticon list to convert each recognised emoticon into a score (+1 or -1) rather than keeping them as separate emoticons, and also by encoding repeated punctuation as the single entity "repeated punctuation" rather than recording each type of repeated punctuation separately. These (language independent) changes would make the machine learning approach more powerful on texts with many emoticons and sentiment-related punctuation, such as the MySpace and Twitter data.

We had difficulty processing the large combined data set, possibly due to the limitations of Weka in terms of processing resources. Initial experiments with a complete set of features needed a computer with large amounts of RAM to load the data. Eventually, 48Gb of RAM (on a 96Gb machine) was assigned to the Java virtual machine but although the data loaded, some of the algorithms ran slowly. For instance Logistic Regression did not complete a single evaluation (out of 30) on 1000 features within two weeks so it was impractical to run full evaluations on the large data set. Instead we used more aggressive initial low-frequency feature reduction and removed all features occurring less than 5 times in the data. Some of the algorithms, including Logistic Regression, were still too slow and so only SMO was used for this data set – the second best performing algorithm overall.

SentiStrength 2 was also assessed using 10-fold cross-validation for the supervised case and also with 30 repetitions.

Corpus statistics

Table 1 shows significant differences in data set sizes. Although BBC and Runners World have similar text sizes, Digg texts are half as big and MySpace texts are under a third as big, with Twitter texts being slightly smaller than MySpace.

Table 1. Text size statistics for each data set.

	Mean chars	Mean Words	Texts
BBC	356.44	62.54	1000
Digg	183.32	31.49	1077
MySpace	101.91	20.08	1041
Runners World	335.42	65.13	1046
Twitter	94.55	15.35	4218
YouTube	91.18	17.12	3407
All six combined	146.05	26.18	11790

Overall sentiment distribution

Figures 1 and 2 report the proportion of different positive and negative sentiment strengths in each data set, according to the average human coded values. From this it can be seen that there are important differences. For example, Runners World and MySpace have a high proportion of positive sentiment in comments (about 80%), whereas Digg and BBC have positive sentiment in under 40% of comments – half as many. Negative comments are rare in MySpace and Twitter (70%, 65% contain

no negativity) but more common in Runners World and Digg (30-40% contain no negativity) and very common in the BBC forums (under 20% contain no negativity). Unusually for sentiment analysis, all the corpora are unbalanced, with highly unequal numbers of members of the different available categories. This makes the task of creating a single, universally effective algorithm more difficult. No pairs of data sets have a similar overall sentiment strength profile although MySpace pairs approximately with Runners World and BBC with Digg.

Note that there are few texts with the maximum positive or negative sentiment strength and so it would be reasonable to collapse the two strongest sentiment classes together but this was not done for consistency with the previous SentiStrength study.

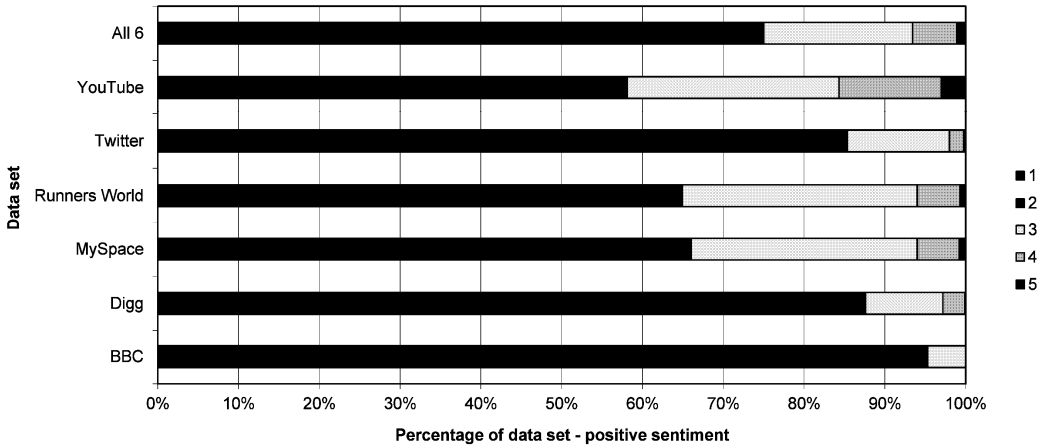

Figure 1. The proportion of positive sentiment strengths in each data set.

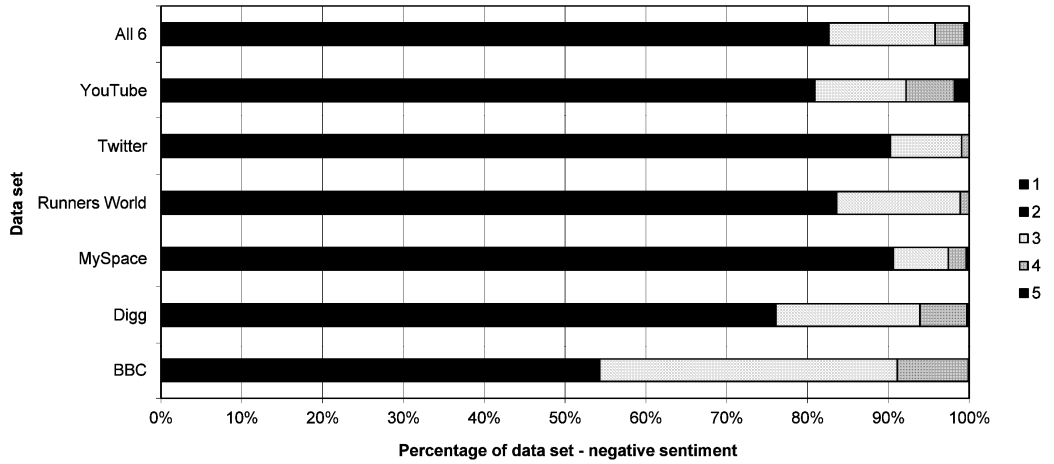

Figure 2. The proportion of negative sentiment strengths in each data set.

Results

From Table 2, SentiStrength exceeds baseline accuracy for negative sentiment strength on all data sets and exceeds baseline accuracy for positive sentiment strength on all data sets except Digg and BBC forums. The most useful measure is correlation because this effectively takes into account the degree of accuracy of each prediction, and so more inaccurate matches get more heavily penalised. A random prediction would get a correlation of 0 and a poor prediction would get a negative correlation but SentiStrength obtains a positive correlation of about 0.3 or higher for all data sets. Hence it is reasonable to use SentiStrength for identifying sentiment patterns in data of any of the types reported in the table. SentiStrength performs weakest in terms of correlation for positive sentiment in Digg and BBC Forums.

Note that whilst supervised SentiStrength tends to be more accurate than unsupervised SentiStrength, they are approximately equal in the key correlation test. This suggests that supervision (i.e., the creation and use of training data to optimise term weights) is not necessary for application domains similar to those in the table.

Table 2. Unsupervised and supervised SentiStrength 2 against the baseline measure (predicting the most common class) and the standard machine learning algorithm and feature set size (from 100, 200 to 1000) having the highest correlation with the human-coded values. Correlation is the most important metric.

BBC Forums*	**+ ve correct**	**- ve correct**	**+ ve +/- 1**	**- ve +/- 1**	**+ ve correl.**	**- ve correl.**
Baseline	63.4%	38.1%	95.3%	91.1%	-	-
Unsupervised ssth	51.3%	46.0%	90.3%	91.1%	0.296	**0.591**
Supervised ssth	60.9% -.2/+.2	48.4% -.3/+.2	94.5% -.1/+.1	92.8% -.1/+.1	0.286 -4/+5	0.573 -3/+2
SLOG 200	**76.7%** -.1/+.1		**97.2%** -0/+.1		**0.508** -4/+4	
SLOG 100		**51.1%** -.2/+.2		**94.7%** -.1/+.1		0.519 -3/+3
Digg						
Baseline	61.5%	46.1%	87.7%	**94.0%**	-	-
Unsupervised ssth	53.9%	46.7%	88.6%	90.8%	0.352	0.552
Supervised ssth	57.9% -.2/+.2	50.5% -.1/+.2	**92.0%** -.1/+.1	92.9% -.1/+.1	**0.380** -3/+3	**0.569** -2/+1
SLOG 100	**63.1%** -.2/+.2		90.9% -.1/+0		0.339 -7/+7	
SLOG 100		**55.2%** -.4/+.3		93.6% -.1/+.2		0.498 -6/+6
MySpace						
Baseline	47.3%	69.9%	94.0%	90.6%	-	-
Unsupervised ssth	62.1%	70.9%	**97.8%**	**95.6%**	**0.647**	0.599
Supervised ssth	62.1% -.3/+.2	72.4% -.1/+.2	96.6% -0/+.1	95.3% -.1/+.1	0.625 -3/+3	**0.615** -2/+3
SLOG 100	**63.0%** -.2/+.2		96.8% -.1/+.1		0.638 -2/+3	
SMO 100		**77.3%** -.1/+.1		93.6% -.1/+.1		0.563 -5/+4
Runners World						
Baseline	44.2%	47.1%	94.0%	**98.9%**	-	-
Unsupervised ssth	53.5%	50.9%	94.7%	90.0%	0.567	0.541
Supervised ssth	53.9% -.3/+.3	55.8% -.3/+.3	**95.4%** -.1/+.1	93.6% -.1/+.1	0.593 -2/+2	0.537 -2/+2
SLOG 200	**61.5%** -.3/+.3		95.3% -.1/+.1		**0.597** -4/+4	
SLOG 300		**65.3%** -.2/+.3		96.1% -.1/+.1		**0.542** -4/+4
Twitter						
Baseline	56.5%	65.7%	85.4%	90.2%	-	-
Unsupervised ssth	59.2%	66.1%	94.2%	93.4%	0.541	0.499
Supervised ssth	63.7% -.1/+0	67.8% -.1/+.1	94.8% -0/+0	94.6% -.1/+0	0.548 -2/+1	0.480 -2/+2
SLOG 200	**70.7%** -.1/+0		**94.9%** -.1/+0		**0.615** -1/+1	
SLOG 200		**75.4%** -.1/+.1		**94.9%** -0/+.1		**0.519** -2/+2
YouTube						
Baseline	31.0%	50.1%	84.3%	80.9%	-	-
Unsupervised ssth	44.3%	56.1%	88.2%	88.5%	0.589	0.521
Supervised ssth	46.5% -.2/+.1	57.8% -.1/+.1	**89.0%** -.1/+0	89.0% -.1/+0	0.621 -1/+1	0.541 -1/+2
SLOG 200	**52.8%** -.1/+.1		89.6% -0/+.1		**0.644** -2/+1	
SLOG 300		**64.3%** -.1/+.1		**90.8%** -.1/+0		**0.573** -3/+3
All 6						
Baseline	42.6%	51.5%	75.1%	82.7%	-	-
Unsupervised ssth	53.5%	58.8%	92.1%	91.5%	0.556	0.565
Supervised ssth	56.3% -0/+.1	61.7% -.1/+.1	**92.6%** -.1/+.1	**93.5%** -0/+0	0.594 -0/+1	**0.573** -1/+0
SMO 800	**60.7%** -0/+.1		92.3% -0/+0		**0.642** -1/+1	
SMO 1000		**64.3%** -0/+.1		92.8% -0/+0		0.547 -1/+2

***** The metrics used are: accuracy (% correct), accuracy within 1 (i.e. +/- 1 class), and correlation. Best values on each data set and each metric are in bold. When multiple tests are available then 30 are conducted and a 95% confidence interval is indicated underneath the mean. For instance, 60.9% above -.2/+.2 denotes a 95% confidence interval for the mean of (60.7%, 61.1%). For correlations, the confidence interval adjustments are for the 3rd decimal place.

Machine learning methods, and logistic regression in particular, tended to be slightly better than SentiStrength. For positive correlations, traditional machine learning performed best on 5 of the 7 datasets, with unsupervised SentiStrength preforming best on the remaining 2. For negative correlations, traditional machine learning performed best on 3 out of 7 datasets, with supervised SentiStrength performing best on 3 and unsupervised SentiStrength performing best on 1. Perhaps surprisingly, however, SentiStrength performed relatively well on the All 6 dataset, despite the large amount of training data. This highlights the domain-dependence of the traditional machine learning approaches, which were presumably not able to fully take advantage of the additional training data because of the multiple domains and genres.

Two data sets for which the machine learning approach performed significantly better than SentiStrength for correlations were the BBC and Twitter positive collections. An investigation into the top features for the BBC revealed many that do not express sentiment. The top 20 were: good, I, hi, “I don’t”, ?, group, 8, be, very, bit, , “a good”, love, “to live in”, thanks, “of your”, “is why”, “I agree”, “the way they”, “by people”, “the field”. Some of these terms clearly express no sentiment but nevertheless associate in the data set with particular positive strengths. The top 100 features list also contains several political terms that were probably the topics of emotional debates rather than used to directly express sentiment, such as “hamas will”, “George Galloway”, “Israel will”, and “that Palestinians”. This shows that the machine learning approach will partly detect emotional topics and not just direct sentiment.

The top 20 features for the Twitter positive data set were: !, http, [any +1 emoticon], lol, love, “I love”, ://bit, “http ://bit”, “[multiple punctuation]! !”, i, “! [multiple punctuation]!”, “! [multiple punctuation]! !”, [#bigrams], [#trigrams], [#words], “[multiple punctuation] !”, good, so, my, you. In this list, square brackets describe a matching feature, quotes are used in multiple term cases (bigrams or trigrams) and all other parts are literal values. The main Twitter features were thus punctuation and length-related as well as parts of URLs (e.g., http://bit.ly URLs). SentiStrength does not incorporate length as part of the algorithm and ignores URLs since they could point to positive or negative content. Presumably, nevertheless, in Twitter people mostly post URLs as recommendations, making positive statements about them. Hence the general machine learning approach is again able to learn from sentiment neutral features to help it to perform better.

Limitations and discussion

A key limitation of the research is that despite the use of six social web data sets with different properties the experiments are not exhaustive and there may still be types of social web environment for which SentiStrength does not work. This seems to be most likely to be the case in environments in which unusual language use is standard, for instance in forums using many jokes or in which sarcasm is widespread.

A second limitation is that not all data sets were coded by three different coders and so the accuracy of the codes for the gold standard may have been weaker on some. This is likely to mean that some of the accuracies reported in Table 2 will be slightly lower than possible, however, and should not affect the answers to the research questions.

Whilst Table 2 gives evidence that it is reasonable to use supervised and unsupervised SentiStrength on a wide variety of social web texts, on most data sets the machine learning approach performed significantly better for overall accuracy and, more importantly, on a small majority it performed better on the key metric of correlation. As the analysis of the results for the high performing machine learning algorithms shows, the machine learning approach can identify and exploit topics that are associated with sentiment (e.g., “George Galloway”, “Israel will”) as well as neutral phrases that nevertheless suggest the presence of sentiment (e.g., “is why”, “of your”). This gives it an advantage that outweighs the knowledge advantage of SentiStrength’s sentiment word list and other rules in some cases and perhaps even in all cases given enough training data. As discussed above, the exploitation of topic is undesirable for some applications, particularly if the focus is on changes in sentiment (Thelwall et al., 2011) or identifying clusters of sentiment (Chmiel et al., 2011),

because the machine learning approach may detect topic changes or topic clusters rather than sentiment changes or sentiment clusters. Moreover, the machine learning approach is more subject to changes over time because topics may change their sentiment association. For example, in the BBC data set, if a peaceful settlement is agreed between Israel and Palestine then these two nouns may become associated with strong positive sentiment, and may also have been in the period when Barack Obama was awarded the 2009 Nobel Peace Prize.

Although all examples discussed so far have been mainly relevant to social science research, exploiting indirect affective terms can also be a problem in some commercial applications. For instance, when designing programs to predict trends using sentiment (Bollen, Pepe, & Mao, 2011) traditional machine learning may predict based upon topic shifts rather than sentiment shifts and could conceivably reduce predictive power over a less accurate approach relying upon direct affective terms. This is based upon the untested assumption that topic shifts would induce more systematic biases than the errors in lower accuracy algorithms using only direct affective terms. Finally, in commercial applications, direct affective terms and transparent methods may be an advantage in contexts where clients see the classified data and naturally wish to understand the reasons for the classifications.

An additional limitation is that SentiStrength does not guarantee to use only direct affective terms because some of the terms in its index are ambiguous, such as *like*, and because even sentiment terms can be used in neutral contexts, as in the case of the word *shocking* in the colour shocking pink. The claim that can be made for SentiStrength is therefore that is has higher reliance upon direct affective terms than machine learning approaches with typical feature sets. Although this has not been directly proven, it seems clear from a comparison of the way in which the two alternative methods work.

Finally, the performance of the machine learning algorithms in Table 2 may be exaggerated because only the best results out of 110 was used in each case (8 algorithms and 3 SVMLight subsumption variations, 10 feature set sizes) except for the combined data set (1 algorithm and 3 SVMLight subsumption variations, 10 feature sets). This is probably not important, however, since the same combination was best in most cases, giving confidence that it is robustly optimal for social web data, at least for a training set of about 1000 texts. It seems that more features than 100 would be optimal for larger training sets, as was the case for the three largest training sets (Twitter, YouTube and combined; Runners World is an anomaly in this context).

Conclusions

The results show that SentiStrength performs significantly above the baseline for correlation across six social web data sets that are substantially different in origin, length and sentiment content. This gives some confidence that SentiStrength is a robust algorithm for sentiment strength detection on social web data. Moreover, this is true for both unsupervised and supervised variants of SentiStrength and so the unsupervised version is a reasonable choice for sentiment strength detection in social web contexts for which no training data is available. This gives positive answers to the first two research questions.

For the third research question, in some environments SentiStrength does not perform as well as some machine learning techniques: particularly logistic regression. Nevertheless, the additional analysis confirmed that the machine learning approach may outperform SentiStrength due to identifying topic or discourse features indirectly associated with sentiment rather than by directly identifying sentiment. As discussed above, this is a problem for some applications.

In conclusion, SentiStrength seems to be suitable for sentiment strength detection in the social web even in its unsupervised version and is recommended for applications in which exploiting only direct affective terms is important. Its major weakness seems to be detecting sarcasm and irony and so this is a logical direction for future research. If reliance upon indirect affective terms is not a problem and sufficient human coded data is available then logistic regression is recommended for social web sentiment strength detection in some contexts, and particularly those with news-related discussions or with significantly more than 1000 human coded training examples. Nevertheless, initial testing suggests that SentiStrength 2 does not perform well on review texts because of the importance of non-sentiment terms like “heavy” and “large” to product review judgements. Finally, in conjunction with previous results on polarity and objectivity detection (Paltoglou & Thelwall, in press; Taboada et al., 2011), there is now a growing body of evidence that sentiment analysis based upon a lexicon and additional rules is broadly robust and relatively domain-independent.

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).

References

Andreevskaia, A., & Bergler, S. (2008). When specialists and generalists work together: Overcoming domain dependence in sentiment tagging. In *Proceedings of ACL-08: HLT* (Vol. Columbus, OH, pp. 290-298): ACL.

Artstein, R., & Poesio, M. (2008). Inter-coder agreement for computational linguistics. *Journal of Computational Linguistics, 34*(4), 555-596.

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.

Balahur, A., Kozareva, Z., & Montoyo, A. (2009). Determining the polarity and source of opinions expressed in political debates *Lecture Notes in Computer Science, 5449*, 468-480.

Balahur, A., Steinberger, R., Kabadjov, M., Zavarella, V., Goot, E. v. d., Halkia, M., et al. (2010). Sentiment analysis in the news. *Proceedings of the Seventh conference on International Language Resources and Evaluation*, Retrieved May 25, 2010 from: http://www.lrec-conf.org/proceedings/lrec2010/pdf/2909_Paper.pdf.

Bollen, J., Gonçalves, B., Ruan, G., & Mao, H. (2011). Happiness is assortative in online social networks. *Artificial Life* , 17(2), 237-251.

Bollen, J., Pepe, A., & Mao, H. (2011). Modeling public mood and emotion: Twitter sentiment and socioeconomic phenomena. *ICWSM11, Barcelona, Spain*, Retrieved June 2, 2011 from: http://arxiv.org/abs/0911.1583.

Bradley, M. M., & Lang, P. J. (1999). Affective norms for English words (ANEW): Instruction manual and affective ratings. Technical Report C-1. *The Center for Research in Psychophysiology, University of Florida,* Retrieved March 6, 2011 from: http://dionysus.psych.wisc.edu/methods/Stim/ANEW/ANEW.pdf.

Chmiel, A., Sienkiewicz, J., Paltoglou, G., Buckley, K., Thelwall, M., & Holyst, J.A. (2011). Negative emotions accelerating users activity in BBC Forum. *Physica A, 390(16), 2936-2944.*

Diakopoulos, N. A., & Shamma, D. A. (2010). Characterizing debate performance via aggregated Twitter sentiment. In *Proceedings of the 28th international conference on Human factors in computing systems* (pp. 1195-1198). New York, NY: ACM Press.

Gamon, M., Aue, A., Corston-Oliver, S., & Ringger, E. (2005). Pulse: Mining customer opinions from free text. *Lecture Notes in Computer Science, 3646*, 121-132.

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.

Hatzivassiloglou, V., & McKeown, K. (1997). Predicting the semantic orientation of adjectives. In *Proceedings of 35th Meeting of the Association for Computational Linguistics* (pp. 174–181). Madrid, Spain.

Joachims, T. (1999). Making large-scale SVM learning practical. In B. Schölkopf, C. Burges & A. Smola (Eds.), *Advances in Kernel Methods - Support Vector Learning* (pp. 41-56): MIT Press.

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.

Krippendorff, K. (2004). *Content analysis: An introduction to its methodology*.Thousand Oaks, CA: Sage.

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

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.

Mishne, G. (2005). Experiments with mood classification in Blog posts. *Style - the 1st Workshop on Stylistic Analysis Of Text For Information Access, at SIGIR 2005*, Retrieved January 10, 2009 from: http://staff.science.uva.nl/~gilad/pubs/style2005-blogmoods.pdf.

Mishne, G., & de Rijke, M. (2006). Capturing global mood levels using Blog posts. In *Proceedings of the AAAI Spring Symposium on Computational Approaches to Analysing Weblogs (AAAI-CAAW)* (pp. 145-152). Menlo Park, CA: AAAI Press.

Mitrovic, M., Paltoglou, G., & Tadic, B. (2011). Networks and emotion-driven user communities at popular Blogs. *The European Physical Journal B, 77*(4), 597-609.

Naveed, N., Gottron, T., Kunegis, J., & Alhadi, A. C. (2011). Bad news travel fast: A content-based analysis of interestingness on Twitter. *WebSci 2011*, Retrieved July 16, 2011 from: http://www.websci2011.org/fileadmin/websci/Papers/2050_paper.pdf.

Neviarouskaya, A., Prendinger, H., & Ishizuka, M. (2007). Textual affect sensing for sociable and expressive online communication. *Lecture Notes in Computer Science, 4738*, 218-229.

Neviarouskaya, A., Prendinger, H., & Ishizuka, M. (2010). Recognition of fine-grained emotions from text: An approach based on the compositionality principle. In T. Nishida, L. Jain & C. Faucher (Eds.), *Modelling machine emotions for realizing intelligence: Foundations and applications* (pp. 179-207).

Neviarouskaya, A., Prendinger, H., & Ishizuka, M. (2011). SentiFul: A lexicon for sentiment analysis. *IEEE Transactions on Affective Computing, 2*(1), 22-36.

O'Connor, B., Balasubramanyan, R., Routledge, B. R., & Smith, N. A. (2010). From tweets to polls: Linking text sentiment to public opinion time series. In *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media* (pp. 122-129). Menlo Park, CA: The AAAI Press

Pak, A., & Paroubek, P. (2010). Twitter as a corpus for sentiment analysis and opinion mining. *Proceedings of LREC 2010*, Retrieved October 5, 2010 from: http://www.lrec-conf.org/proceedings/lrec2010/pdf/2385_Paper.pdf.

Paltoglou, G., & Thelwall, M. (in press). Twitter, MySpace, Digg: Unsupervised sentiment analysis in social media. *ACM Transactions on Intelligent Systems and Technology*, Abstract available at: http://tist.acm.org/papers/TIST-2010-2011-0317.html.

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

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

Polanyi, L., & Zaenen, A. (2006). Contextual valence shifters. In J. Wiebe (Ed.), *Computing attitude and affect in text: Theory and applications* (pp. 1-10). Dordrecht: Springer.

Read, J. (2005). Using emoticons to reduce dependency in machine learning techniques for sentiment classification. *Proceedings of the ACL 2005 Student Research Workshop*, 43-48.

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., & 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.

Strapparava, C., & Valitutti, A. (2004). WordNet-affect: an affective extension of WordNet. In *Proceedings of the 4th International Conference on Language Resources and Evaluation* (pp. 1083-1086). Lisbon.

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.

Taboada, M., Brooke, J., & Stede, M. (2009). Genre-based paragraph classification for sentiment analysis. In Proceedings of SIGDIAL 2009: The 10th Annual Meeting of the Special Interest

Group in Discourse and Dialogue (pp. 62–70). London: Association for Computational Linguistics.

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

Taboada, M., & Grieve, J. (2004). Analyzing appraisal automatically. In *Proceedings of AAAI Spring Symposium on Exploring Attitude and Affect in Text (AAAI Technical Report SS-04-07)* (pp. 158–161). Stanford University, CA: AAAI Press.

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.

Tang, H., Tan, S., & Cheng, X. (2009). A survey on sentiment detection of reviews *Expert Systems with Applications: An International Journal, 36*(7), 10760-10773.

Thelwall, M. (2010). Emotion homophily in social network site messages. *First Monday, 10*(4), Retrieved March 6, 2011 from: http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/2897/2483.

Thelwall, M., Buckley, K., & Paltoglou, G. (2011). Sentiment in Twitter events. *Journal of the American Society for Information Science and Technology, 62*(2), 406-418.

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.

Thelwall, M., Wilkinson, D., & Uppal, S. (2010). Data mining emotion in social network communication: Gender differences in MySpace. *Journal of the American Society for Information Science and Technology, 21*(1), 190-199.

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.

Velikovich, L., Blair-Goldensohn, S., Hannan, K., & McDonald, R. (2010). The viability of web-derived polarity lexicons. *Proceedings of the 11th Conference of the North American Association for Computational Linguistics*, Retrieved June 11, 2011 from: http://www.ryanmcd.com/papers/web_polarity_lexiconsNAACL2010.pdf.

Wilson, T., Wiebe, J., & Hoffman, P. (2009). Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis. *Computational linguistics, 35*(3), 399-433.

Wilson, T., Wiebe, J., & Hwa, R. (2006). Recognizing strong and weak opinion clauses. *Computational Intelligence, 22*(2), 73-99.

Witten, I. H., & Frank, E. (2005). *Data mining: Practical machine learning tools and techniques*. San Francisco: Morgan Kaufmann.

Wu, Q., Tan, S., Duan, M., & Cheng, X. (2010). A two-stage algorithm for domain adaptation with application to sentiment transfer problems. *Lecture Notes in Computer Science, 6458*, 443-453.

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