Sentiment Strength Detection in Short Informal Text[1]

Mike Thelwall, Kevan Buckley, Georgios Paltoglou, Di Cai
Statistical Cybermetrics Research Group, School of Computing and Information 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, caid@wlv.ac.uk
Tel: +44 1902 321470 Fax: +44 1902 321478
Arvid Kappas
School of Humanities and Social Sciences, Jacobs University Bremen, Campus Ring 1, 28759 Bremen, Germany
E-mail: a.kappas@jacobs-university.de
Tel: +49 421 200-3441

A huge number of informal messages are posted every day in social network sites, blogs and discussion forums. Emotions seem to be frequently important in these texts for expressing friendship, showing social support or as part of online arguments. Algorithms to identify sentiment and sentiment strength are needed to help understand the role of emotion in this informal communication and also to identify inappropriate or anomalous affective utterances, potentially associated with threatening behaviour to the self or others. Nevertheless, existing sentiment detection algorithms tend to be commercially-oriented, designed to identify opinions about products rather than user behaviours. This article partly fills this gap with a new algorithm, SentiStrength, to extract sentiment strength from informal English text, using new methods to exploit the de-facto grammars and spelling styles of cyberspace. Applied to MySpace comments and with a lookup table of term sentiment strengths optimised by machine learning, SentiStrength is able to predict positive emotion with 60.6% accuracy and negative emotion with 72.8% accuracy, both based upon strength scales of 1-5. The former, but not the latter, is better than baseline and a wide range of general machine learning approaches.

Introduction

Most opinion mining algorithms attempt to identify the polarity of sentiment in text: positive, negative or neutral. Whilst for many applications this is sufficient, texts often contain a mix of positive and negative sentiment and for some applications it is necessary to detect both simultaneously and also to detect the strength of sentiment expressed. For instance, programs to monitor sentiment in online communication, perhaps designed to identify and intervene when inappropriate emotions are used or to identify at-risk users (e.g., Huang, Goh, & Liew, 2007), would need to be sensitive to the strength of sentiment expressed and whether participants were appropriately balancing positive and negative sentiment. In addition, basic research to understand the role of emotion in online communication (e.g., Derks, Fischer, & Bos, 2008; e.g., Hancock, Gee, Ciaccio, & Lin, 2008; Nardi, 2005) would also benefit from fine-grained sentiment detection, as would the growing body of psychology and other social science research into the role of sentiment in various types of discussion or general discourse (Balahur, Kozareva, & Montoyo, 2009; Pennebaker, Mehl, & Niederhoffer, 2003; Short & Palmer, 2008).

A complicating factor for online sentiment detection is that there are many electronic communications media in which text based communication in English seems to frequently ignore the rules of grammar and spelling. Perhaps most famous is mobile phone text language with its abbreviations, emoticons and truncated sentences (Grinter & Eldridge, 2003; Thurlow, 2003) but similar styles are evident in many other forms of computer mediated

[1] 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.

communication, including chatrooms, bulletin boards and social network sites (Baron, 2003; Crystal, 2006). Widely recognised innovations include emoticons like :-) that are reasonably effective in conveying emotion (Derks, Bos, & von Grumbkow, 2008; Fullwood & Martino, 2007) and word abbreviations like m8 (mate) and u (you) (Thurlow, 2003). Although sometimes seen as poor language use, these are a natural response to the technological affordances and social factors associated with a system (Baron, 2003; Walther & Parks, 2002). These variations cause problems because typical linguistic sentiment analysis programs start with part of speech tagging (e.g., Brill, 1992), which is reliant upon standard spelling and grammar, and/or apply rules that assume at least correct spelling, if not correct grammar. Spelling correction can be useful in this context, but this is based upon the assumption that spelling deviations are likely to be accidental mistakes (Kukich, 1992; Pollock & Zamora, 1984) and so current algorithms are unlikely to work well with deliberately non-standard spellings. Nevertheless, there is a range of common abbreviations and new words that a linguistic algorithm could, in principle, detect. Non-linguistic machine learning algorithms typically predict sentiment based upon occurrences of individual words, word pairs and word triples in documents. These may also perform poorly on informal text because of spelling problems and creativity in sentiment expression, even if a large training corpus is available (see below).

The social network site MySpace, the source of the data used in the current study, is known for its young members, its musical orientation and its informal communication patterns (boyd, 2008; boyd, 2008). Probably as a result of these factors 95% of English public comments exchanged between friends contain at least one abbreviation from standard English (Thelwall, 2009). Common features include emoticons, texting-style abbreviations and the use of repeated letters or punctuation for emphasis (e.g., a loooong time, Hi!!!). Comments are typically short (mean 18.7 words, median 13 words, 68 characters) (Thelwall, 2009) but positive emotion is common (Thelwall, Wilkinson, & Uppal, 2010).

This article proposes a new algorithm, SentiStrength, which employs several novel methods to simultaneously extract positive and negative sentiment strength from short informal electronic text. SentiStrength uses a dictionary of sentiment words with associated strength measures and exploits a range of recognised non-standard spellings and other common textual methods of expressing sentiment. SentiStrength was developed through an initial set of 2,600 human-classified MySpace comments, and evaluated on a further random sample of 1,041 MySpace comments. Note that in some articles, but not in emotion psychology, the term sentiment refers to affect split into positive, negative and neutral whereas the term emotion refers to more differentiated affect (e.g., happy, sad, frightened). In contrast, the two terms are used as synonyms here, with their meaning effectively defined by the coder instructions described below. The main novel contributions of this paper are: a machine learning approach to optimise sentiment term weightings; methods for extracting sentiment from repeated letter non-standard spelling in informal text; and a related spelling correction method. In addition, the paper introduces a dual 5-point system for positive and negative sentiment, a corpus of 1,041 MySpace comments for this system, and a new overall sentiment strength detection system that combines novel and existing methods.

Background and Related Work

This literature review section discussed related opinion mining/sentiment analysis research as well as some relevant contributions from emotion psychology.

Opinion mining

Opinion mining, also known as sentiment analysis, is the extraction of positive or negative opinions from (unstructured) text (Pang & Lee, 2008). The many applications of opinion mining include detecting movie popularity from multiple online reviews and diagnosing which parts of a vehicle are liked or disliked by owners through their comments in a dedicated site or forum. There are also applications unrelated to marketing, such as

differentiating between emotional and informative social media content (Denecke & Nejdl, 2009).

Opinion mining typically occurs in two or three stages, although more may be needed for some tasks (e.g., Balahur et al., 2010). First, the input text is split into sections, such as sentences, and each section tested to see if it contains any sentiment: if it is subjective or objective (Pang & Lee, 2004). Second, the subjective sentences are analysed to detect their sentiment polarity. Finally, the object about which the opinion is expressed may be extracted (e.g., Gamon, Aue, Corston-Oliver, & Ringger, 2005). Opinion mining normally deals with only positive and negative sentiment rather than discrete emotions (e.g., happiness, surprise), does not detect sentiment strength (but sometimes uses the strength of association of words with positive or negative sentiment, e.g., Kaji & Kitsuregawa, 2007), and does not simultaneously identify both positive and negative emotions. Nevertheless, such opinion mining research can aid the simultaneous assessment of positive and negative sentiment strength both because of its general insights into sentiment analysis and also because most techniques could, in theory, be repurposed for this new task. For example, phrase analysis techniques could be applied to identify both positive and negative sentiment even within individual sentences (Choi & Cardie, 2008; Wilson, 2008; Wilson, Wiebe, & Hoffman, 2009).

Opinion mining algorithms often use machine learning to identify general features associated with positive and negative sentiment, where these features could be a subset of the words in the document, parts of speech or n-grams (i.e., the frequency of occurrence of all n consecutive words, where n is typically 1, 2, or 3) (Abbasi, Chen, Thoms, & Fu, 2008; Ng, Dasgupta, & Arifin, 2006; Tang, Tan, & Cheng, 2009). Other features used with some success include: emoticons in online movie reviews (Read, 2005), which seem so be more domain-independent than words; lexico-syntactic patterns (e.g., Riloff & Wiebe, 2003); and artificial features derived from adjective polarity lists (Ng et al., 2006). The additional features typically provide small but significant increases in performance. Rules-based methods have also been used to identify structures in sentences associated with sentiment (Prabowo & Thelwall, 2009; Wu, Chuang, & Lin, 2006). Two recurring machine learning issues are *feature selection* and *classification algorithm choice*.

Feature selection, data processing to remove the least useful n-grams, has been shown to slightly improve classification performance, for example by choosing a restricted set of features (e.g., 5000) that score highest on a measure like information gain (Riloff, Patwardhan, & Wiebe, 2006), or log likelihood (Gamon, 2004). When using n-grams (and lexico-syntactic patterns) small improvements can also be made by pruning the feature set of features that are subsumed by simpler features that have stronger information gain values (Riloff et al., 2006). For example, if "love" has a much higher information gain value than "I love" then the bigram can be eliminated without much risk of loss of power for the subsequent classification. An entropy-weighted genetic algorithm can also perform better than standard feature reduction approaches (Abbasi, Chen, & Salem, 2008).

In terms of classification algorithms, support vector machines (SVMs) are widely used (Abbasi et al., 2008; Abbasi et al., 2008; Argamon et al., 2007; Gamon, 2004; Mishne, 2005; Wilson, Wiebe, & Hwa, 2006) because they seem to perform as well or better than other methods in most machine learning contexts. Nevertheless, with a few exceptions (Read, 2005; Wilson et al., 2006), explicit comparisons with other methods have not been included in opinion mining publications.

Many other approaches have also been used to detect sentiment in text. One is to have a dictionary of positive and negative words (e.g., love, hate), such as that found in General Inquirer (Stone, Dunphy, Smith, & Ogilvie, 1966), WordNet Affect (Strapparava & Valitutti, 2004), SentiWordNet (Baccianella, Esuli, & Sebastiani, 2010; Esuli & Sebastiani, 2006) or Q-WordNet (Agerri & García-Serrano, 2010), and to count how often they occur. Modifications of this approach include the identification of negating terms (Das & Chen, 2001), words that enhance sentiment in other words (e.g., *really* love, *absolutely* hate) and overall sentence structures (Turney, 2002). A more sophisticated approach is to identify text features that could potentially be subjective in some contexts and then use contextual information to decide

whether they are subjective in each new context (Wiebe, Wilson, Bruce, Bell, & Martin, 2004).

An alternative opinion mining technique has used a primarily linguistic approach: simple rules based upon compositional semantics (information about likely meanings of a word based upon the surrounding text) to detect the polarity of an expression (Choi & Cardie, 2008). This gives good results on phrases in newswire documents that are manually coded as having at least medium level positive or negative sentiment. This approach seems particularly suited to cases where there is a large volume of grammatically correct text from which rules can be learned. Nevertheless, a study of poor grammatical quality texts in online customer feedback showed that linguistic approaches *could* improve classification slightly when added to bag of words (1-grams) approaches, although aggressive feature reduction had a similar impact to adding linguistic features (Gamon, 2004). The improvement was probably due to the large data set available (40,884 documents with an average of 2.26 sentences each), as has been previously claimed for an analysis of informal text (Mishne, 2005). Another approach used a lexicon of appraisal adjectives (e.g., "sort of", "very") together with an orientation lexicon to detect movie review polarity. This did not perform as well as unigrams but the combined performance was better than that of unigrams alone (Argamon et al., 2007). Linguistic features have also been successfully used to extend opinion mining to a multi-aspect variant that is able to detect opinions about different aspects of a topic (Snyder & Barzilay, 2007). A promising future approach is the incorporation of context about the reasons why sentiment is used, such as differentiating between intention, arguments and speculation (Wilson, 2008).

Detecting multiple emotions

Psychology of emotion research argues that whilst positive and negative sentiment are important dimensions, there are many different widely socially-recognised types of emotion and the strength of emotions (arousal level) can vary (e.g., Cornelius, 1996; Fox, 2008). In the dimensional model of emotion from psychology (Russell, 1979), sentiment can always be fundamentally split into two axes: arousal (low to high) and valence (positive to negative). Whilst this model is useful, other research has shown that positive and negative sentiment can coexist (e.g., Fox, 2008, p. 127) and are relatively independent in many contexts – particularly when sentiment levels are not extreme and over longer time periods (Diener & Emmons, 1984; Huppert & Whittington, 2003; Watson, 1988; Watson, Clark, & Tellegen, 1988) and so it also seems reasonable to conceive sentiment as separately-measureable positive and negative components, as encoded in a popular psychology research instrument (Watson et al., 1988).

There have been some previous attempts to develop algorithms to detect the strength or prevalence of sentiment or emotion in text, or to differentiate between several types of emotion. The LIWC (Linguistic Inquiry and Word Count, www.liwc.net) software from psychology, for example, uses a list of emotion-bearing words to detect positive and negative emotion in text in addition to three specific emotions of particular use in psychology and psychotherapy: anger, anxiety and sadness. It uses simple word counting, measuring the proportion of words falling within an extensive predefined list (e.g., 408 positive and 499 negative words or word stems). The list includes some words that are associated with emotions but do not describe them. For example 'lucky' is a positive keyword and 'loses' is a negative keyword. In contrast to the machine learning approaches discussed above, these lists have been compiled and validated using panels of human judges and statistical testing.

LIWC calculates the *prevalence* of emotion in text, rather than attempting to diagnose a text's overall emotion or emotion strength. It is most suited to longer documents, for which its statistics would be useful indicators of the tendency for emotion to occur. The program uses word truncation for simplicity (e.g., joy* matches any word starting with joy), rather than stemming or lemmatisation, but does not take into account booster words like "very" or the negating effect of negatives (e.g., *not* happy). LIWC has been used by psychology researchers to investigate the connection between language and psychology (Pennebaker et

al., 2003) and also as a practical tool, for example to detect how well people are likely to cope with bereavement based upon their language use (Pennebaker, Mayne, & Francis, 1997). A related emotion detection approach differentiates between happy, unhappy and neutral states based upon words used by students describing their daily lives (Wu et al., 2006). This is similar to the typical positive/negative/neutral objective for opinion mining, however.

One computer science initiative has attempted to identify various emotions in text, focussing on the six so-called basic emotions (Ekman, 1992; Fox, 2008) of anger, disgust, fear, joy, sadness and surprise (Strapparava & Mihalcea, 2008). This initiative also measured emotion strength. A human-annotated corpus was used with the coders allocating a strength from 0 to 100 for each emotion to each text (a news headline), although inter-annotator agreement was low (Pearson correlations of 0.36 to 0.68, depending on the emotion). A variety of algorithms were subsequently trained on this data set. For example, one used WordNet Affect lists to generate appropriate dictionaries for the six emotions. A second approach used a Naive Bayes classifier trained on sets of LiveJournal blogs annotated by their owners with one of the six emotions. The best system (for fine-grained evaluation) was one previously designed for newspaper headlines, UPAR7 (Chaumartin, 2007), which used linguistic parsing and tagging as well as WordNet, SentiWordNet and WordNet Affect, hence relying upon reasonably correct standard grammar and spelling.

In psychology, the term mood refers to medium and long term affective states. Some blogs and social network sites allow members to describe their mood at the time of editing their status or writing a post, typically by selecting from a range of icons. The results can be used as annotated mood corpora. In theory such corpora ought to be usable to train classifiers to identify mood from the text associated with the mood icon and one system has been designed to do this, but with limited success, probably because the texts analysed are typically short (average 200 words) and there are many moods, some of which are very similar to each other, although even a binary categorisation task also had limited success (Mishne, 2005). A follow up project attempted to derive the proportion of posts with a given mood within a specific time period using 199 words (1-grams) and word pairs (2-grams) derived from the aggregate of all texts, rather than by classifying individual texts (Mishne & de Rijke, 2006). The results showed a high correlation with aggregate self-reported mood. A similar aggregation approach has been applied subsequently in a range of social science contexts (Hopkins & King, 2010).

Linguistic processing has also been combined with a pre-existing large collection of subjective common sense statement patterns and applied to relatively informal and domain-independent text in email messages to detect multiple emotions (Liu, Lieberman, & Selker, 2003). This was part of an email support system, however, and the accuracy of the emotion detection was not directly evaluated.

Sentiment strength detection

In addition to the research discussed above concerning strength detection for multiple emotions (Strapparava & Mihalcea, 2008), there is some work on positive-negative sentiment strength detection. One previous study used modified sentiment analysis techniques to predict the strength of human ratings on a scale of 1 to 5 for movie reviews (Pang & Lee, 2005). This is a kind of sentiment strength evaluation with a combined scale for positive and negative sentiment. Experiments with human judgements led the authors to merge two of the categories and so the final task was a 4 category classification, with a 3 category version also constructed for testing purposes. A comparison of multi-class SVM classification with SVM regression suggested that SVM regression worked slightly better than multi-class SVM classification when all 4 categories were used but not when only 3 categories were used. It seems likely that the relative performance of SVM regression would increase further as the number of categories increases because the ordering of the classes is implicit information that the multi-class SVM does not use but that SVM regression does. Slight improvements were also gained when information about the percentage of positive sentences in each review was added. This may not be relevant to corpora of very short texts, however.

Sentiment strength classification has also been developed for a three level scheme (low, medium, and high or extreme) for subjective sentences or clauses in newswire texts using a linguistic analysis converting sentences into dependency trees reflecting their structure (Wilson et al., 2006). Adding dependency trees to unigrams substantially improved the performance of various classifiers compared to unigrams alone, perhaps helped by the fairly large training set (9,313 sentences), the (presumably) good quality grammar of the texts, and the fairly low initial performance on this task (34.5% to 50.9% for unigrams, rising to 48.3% to 55.0% for the three types of classifier applied to level 1 clauses). Here, SVM regression was outperformed by both the rule-based learning Ripper (Cohen, 1995) and BoosTexter, a boosting algorithm combining multiple weak classifiers (Schapire & Singer, 2000).

Quite similar to the current paper is one that measured multiple emotions and their strengths in informal text associated with a dialog system using a combination of methods, including seeking symbolic cues via repeated punctuation (e.g., !!), emoticons and capital letters as well as translating abbreviations (Neviarouskaya, Prendinger, & Ishizuka, 2007). The system also measured emotion intensity on a scale of 0-1 and used a dictionary of terms and intensity ratings assigned by three human judges (with moderate agreement rates: Fleiss Kappa 0.58). The reported evaluation on 160 human-coded sentences showed that in 68% of sentences the system agreed with the coder average to within 20%.

Data Set and Human Judgement of Sentiment Strength

MySpace was chosen as a source of test data for this study because it is a public environment containing a large quantity of informal text language and is important in its own right as one of the most visited web sites in the world in 2009. A random sample of MySpace comments was taken by examining the profiles of every 15th member that joined on June 18, 2007, up to 40,000 and selecting those with a declared U.S. nationality and a public profile not of a musician, comedian or film-maker. Of these, those with less than two friends or no comments were rejected as inactive and those with over 1,000 friends or 4,000 comments were rejected as abnormal. A commenting friend was then identified for each remaining member, satisfying the same criteria above, and a random comment selected from each direction of communication between the two. The comments were extracted in December 2008. This produced a large essentially random sample of U.S. commenter-commentee messages. Spam comments and chain messages were subsequently eliminated, as were comments containing images.

Although sentiment analysis is normally concerned with opinions (Pang & Lee, 2008), Wilson (2008) has generalised this to the psychological task of identifying the author's hidden internal state from their text. For the MySpace data, the objective was not to determine opinions or the author's internal state, however, but to identify the role of expressed sentiment for online communication. Hence the focus of the task was to identify the sentiment expressed in each message, whether reflecting the author's hidden internal state, the intended message interpretation, or the reader's hidden internal state.

In order to obtain reliable human judgements of a random sample of the MySpace comments, two pilot exercises were undertaken with separate samples of the data (a total of 2,600 comments). These were used to identify key judgement issues and an appropriate scale. Although there are many ways to measure emotion (Mauss & Robinson, 2009; Wiebe, Wilson, & Cardie, 2005), human coder subjective judgements were used as an appropriate way to gather sufficient results. A set of coder instructions was drafted and refined and an online system constructed to randomly select comments and present them to the coders. One of the key outcomes from the pilot exercise was that the coders treated expressions of energy as expressions of positive sentiment unless in an explicitly negative context. For example, “Hey!!!” would be interpreted as positive because it expresses energy in a context that gives no clue as to the polarity of the emotion, so it would be accepted by most coders as positive by default. In contrast, “Loser!!!” would be interpreted as more negative than “Loser” as the exclamation marks are associated with a negative word. Consequently, the instructions were

revised to explicitly state that this conflation of ostensibly neutral energy and positive sentiment was permissible.

For the final judgements, over a thousand MySpace comments in the data set (20 words and 101 characters per comment, on average) were selected to be judged on a 5 point scale as follows for both positive and negative sentiment.

[no positive emotion or energy] **1– 2 – 3 – 4 – 5** [very strong positive emotion]

[no negative emotion] **1– 2 – 3 – 4 – 5** [very strong negative emotion]

The coders were given verbal instructions for coding each text as well as a booklet explaining the task (motivated by Wiebe et al., 2005), with the key instructions reproduced in this article's appendix. The booklet also contained a list of emoticons and acronyms with explanations and background context of the task for motivation purposes. An early version of the booklet included examples of comments with associated positive and negative sentiment judgements but these had little impact in practice on coders during the pilot testing phase. The set of examples was therefore not used so that inter-coder reliability could be more realistically assessed without the possibility that some of the comments were too similar to the examples given.

Emotions are perceived differently by individuals, partly because of their life experiences and partly because of personality issues (Barrett, 2006) and gender (Stoppard & Gunn Gruchy, 1993). For system development, the judgements should give a consistent perspective on sentiment in the data, rather than an estimate of the population average perception. As a result, a set of same gender (female) coders was used and initial testing conducted to identify a homogeneous subset. Five coders were initially selected but two were subsequently rejected for giving anomalous results: one gave much higher positive scores than the others, and another gave generally inconsistent results. The mean of the three coders' results was calculated for each comment and rounded. This was the gold standard for the experiments. Below are some examples of texts and judgements.

- hey witch wat cha been up too (scores: +ve: 2,3,1; -ve: 2,2,2)
- omg my son has the same b-day as you lol (scores: +ve: 4,3,1; -ve: 1,1,1)
- HEY U HAVE TWO FRIENDS!! (scores: +ve: 2,3,2; -ve: 1,1,1)
- What's up with that boy Carson? (scores: +ve: 1,1,1; -ve: 3,2,1)

Table 1 reports the degree of inter-coder agreement. Basic agreement rates are reported here for comparability with SentiStrength. Previous emotion-judgement/annotation tasks have obtained higher inter-coder scores, but without strength measures and therefore having fewer categories (e.g., Wiebe et al., 2005). Moreover, one previous paper noted that inter-coder agreement was higher on longer (blog) texts (Gill, Gergle, French, & Oberlander, 2008), suggesting that obtaining agreement on the short texts here would be difficult. The appropriate type of inter-coder reliability statistic for this kind of data with multiple coders and varying differences between categories is Krippendorff's α (Artstein & Poesio, 2008; Krippendorff, 2004). Using the numerical difference in emotion score as weights, the three coder α values were 0.5743 for positive and 0.5634 for negative sentiment. These values are positive enough to indicate that there is broad agreement between the coders but not positive enough (e.g., < 0.67. although precise limits are not applicable to Krippendorff's α with weights) to suggest that the coders are consistently measuring a clear underlying construct. Nevertheless, using the *average* of the coders as the gold standard still seems to be a reasonable method to get sentiment strength estimates.

Table 1. Level of agreement between coders for the 1,041 evaluation comments (exact agreement, % of agreements within one class, mean percentage error, and Pearson correlation).

Comparison	**+ve**	**+ve +/- 1 class**	**+ve mean % diff.**	**+ve corr**	**-ve**	**-ve +/- 1 class**	**-ve mean % diff.**	**-ve corr**
Coder 1 vs. 2	51.0%	94.3%	.256	.564	67.3%	94.2%	.208	.643
Coder 1 vs. 3	55.7%	97.8%	.216	.677	76.3%	95.8%	.149	.664
Coder 2 vs. 3	61.4%	95.2%	.199	.682	68.2%	93.6%	.206	.639

The SentiStrength Sentiment Strength Detection Algorithm

The SentiStrength emotion detection algorithm was developed on an initial set of 2,600 MySpace classifications used for the pilot testing. The key elements of SentiStrength are listed below.

- The core of the algorithm is the **sentiment word strength list.** This is a collection of 298 positive terms and 465 negative terms classified for either positive or negative sentiment strength with a value from 2 to 5. The default classifications are based upon human judgements during the development stage, with automatic modification occurring later during the training phase (see below). Following LIWC, some of the words include wild cards (e.g., xx*) matches any number ≥2 of consecutive xs. Some terms are standard English words and others are non-standard but common in MySpace (e.g., luv, xox, lol, haha, muah). The emotion strength is specific to the contexts in which the words tend to be used in MySpace. For example, "love" was originally classified as strength 4 positive but was reduced to strength 3 due to many casual uses such as "Just showin love 2 ur page". Some of the words explicitly express emotion, such as "love" or "hate" but others, normally given a weak strength 2, are indirectly associated with positive or negative contexts (e.g., appreciate, help, birthday). The SentiStrength algorithm includes procedures (described below) to fine-tune the sentiment strengths using a set of training data.
- The above default manual word strengths are modified by a **training algorithm to optimise the sentiment word strengths**. This algorithms starts with the baseline human-allocated term strengths for the predefined list and then for each term assesses whether an increase or decrease of the strength by 1 would increase the accuracy of the classifications. Any change that increases the overall accuracy by at least 2 is kept. The minimum increase could also be set to 1 which would risk over-fitting, whereas 2 risks loosing useful changes to rare words. Here 2 was selected to make the algorithm run faster, due to less changes, rather than for any theoretical reason (in fact the algorithm worked better on the test data with 1, as the results show). The algorithm tests all words in the sentiment list at random and is repeated until all words have been checked without their strengths being changed.
- The word "**miss**" was allocated a positive and negative strength of 2. This was the only word classed as both positive and negative. It was typically used in the phrase "I miss you", suggesting both sadness and love.
- A **spelling correction algorithm** identifies the standard spellings of words that have been miss-spelled by the inclusion of repeated letters. For example hellllloooo would be identified as "hello" by this algorithm. The algorithm (a) automatically deletes repeated letters above twice (e.g., helllo -> hello); (b) deletes repeated letters occurring twice for letters rarely occurring twice in English (e.g., niice -> nice) and (c) deletes letters occurring twice if not a standard word but would form a standard word if deleted (e.g., nnice -> nice but not hoop -> hop nor baaz -> baz). Formal spelling correction algorithms (see Pollock & Zamora, 1984) were tried but not used as they made very few corrections and had problems with names and slang.
- A **booster word list** contains words that boost or reduce the emotion of subsequent words, whether positive or negative. Each word increases emotion strength by 1 or 2 (e.g., very, extremely) or decreases it by 1 (e.g., some).
- A **negating word list** contains words that invert subsequent emotion words (including any preceding booster words). For example, if "very happy" had positive strength 4 then "not very happy" would have negative strength 4. The possibility that some negating terms do not negate was not incorporated as this did not seem to occur often in the pilot data set.
- **Repeated letters** above those needed for correct spelling are used to give a strength boost of 1 to emotion words, as long as there are at least two additional letters. The use of repeated letters is a common device for expressing emotion or energy in MySpace comments, but one repeated letter often appeared to be a typing error.

- An **emoticon list** with associated strengths (positive or negative 2) supplements the sentiment word strength list (and punctuation included in emoticons is not processed further for the purposes below).
- Any sentence with an **exclamation mark** was allocated a minimum positive strength of 2.
- **Repeated punctuation** including at least one exclamation mark gives a strength boost of 1 to the immediately preceding emotion word (or sentence).
- **Negative emotion was ignored in questions**. For example, the question "are you angry?" would be classified as not containing sentiment, despite the presence of the word "angry". This was not applied to positive sentiment because many question sentences appeared to contain mild positive sentiment. In particular, sentences like "whats up?" were typically classified as containing mild positive sentiment (strength 2).

The above factors were applied separately to each sentence, with the sentence being assigned with both the most positive and most negative emotion identified in it. Each overall comment was assigned with the most positive of its sentence emotions and the most negative of its sentence emotions. Sentence were split either by line breaks in comments or after punctuation other than emoticons.

Some additional modifications were added to SentiStrength but subsequently rejected after additional testing, or were found to be impractical.

- Phrase identification was *not* extensively used except for a few frequent examples found in the initial 2,600 development comments. Although idiomatic phrases were common, their variety was such that it did not seem practical to systematically identify them. Future work could perhaps identify booster phrases like "so much" and "a lot", and use phrase identification to separate weak uses of the word "love" with stronger uses, such as "I love you".
- Semantic disambiguation was *not* used for ambiguous words because of the problems caused by highly non-standard grammar. This could potentially improve the algorithm but would require considerable computational effort. For example, the word "rock" was sometimes strongly positive (e.g., you rock!!!) and sometime neutral (e.g., do you listen to rock music?).

Experiments

SentiStrength was tested on a set of 1,041 MySpace comments that were different from the comments used in the development phase and were classified by three people (see Table 1), and the average was used as the gold standard. A 10-fold cross-validation approach was used. The results were compared to random allocation and to the baseline majority class classification (a positive sentiment of 2 and a negative sentiment of 1). SentiStrength was also compared to a range of standard machine learning classification algorithms in Weka (Witten & Frank, 2005) using the frequencies of each word in the sentiment word list as the feature set. The *extended feature set* used for the comparisons included n-grams of length 1-3 consisting of all terms extracted from the text, including emoticons, spelling-corrected words (where appropriate), repeated punctuation, question marks and exclamation marks (e.g., one feature was the 3-gram: "love-u-!") as well as counts of the total number of 1, 2, and 3-grams in each comment. This extended set of features incorporates most of the elements of text used by SentiStrength.

A second test compared different feature sets to see whether alternative smaller feature sets could give better results for machine learning and to discover which features were most useful.

A third test used feature reduction with subsumption (see below for details).

A fourth test compared different variations of SentiStrength to see which aspects of the algorithm were most powerful.

Comparison with machine learning, extended feature set

Figures 1 and 2 show the performance of various machine learning algorithms on the 1,041 MySpace comments with different feature set sizes, as selected using the top-ranking features from the information gain metric. Feature selection improved the results for all methods, with one minor exception (Naïve Bayes for positive sentiment: 52.0% without feature selection, averaged over 4 10-fold cross-validations). For each method, Table 2 reports comparisons with SentiStrength using the optimal feature set size for each method.

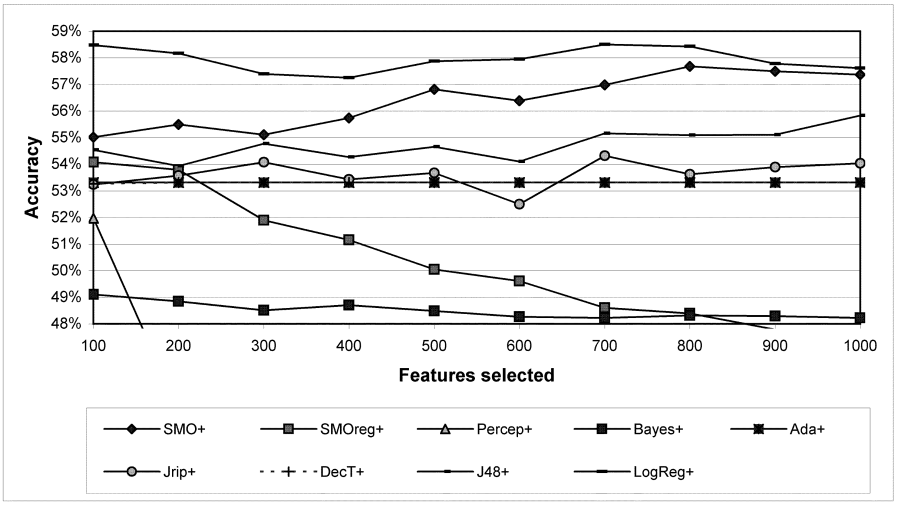

Fig. 1. Positive sentiment classification accuracy against feature set size for different classifiers using the extended feature set; average over 4 classifications.

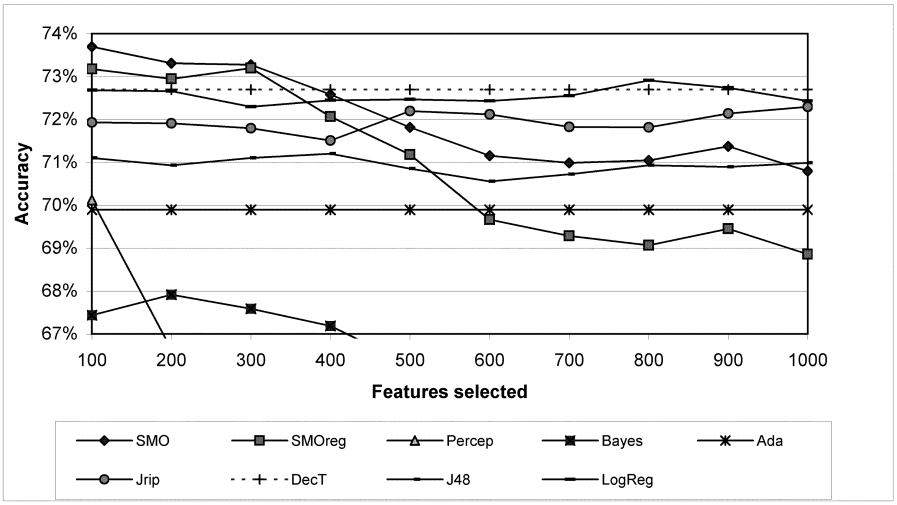

Fig. 2. Negative sentiment classification accuracy against feature set size for different classifiers using the extended feature set; average over 4 classifications.

From Table 2, machine learning classifiers using the extended feature set with the optimal number of features, as selected by information gain, are significantly less accurate than SentiStrength. SentiStrength also has the highest correlation with the gold standard, the lowest mean percentage error and the highest accuracy to within one class. Hence it performs consistently better (at least 2.1%) than the other algorithms. The level of accuracy for SentiStrength is nevertheless moderate at 60.6%. This is similar to the degree of agreement between the human coders (Table 1), suggesting that positive sentiment strength detection in informal short texts is an inherently difficult task.

Table 2. Performance of various algorithms on positive sentiment strength detection for 1,041 comments with the extended feature set and 10-fold cross-validation (decreasing order of positive sentiment strength performance). Other than SentiStrength, results are averages over 4 runs of different random test/training splits and for the optimal feature numbers, as selected from Figure 1.

Algorithm	**Optimal features**	**Accuracy**	**Accuracy +/- 1 class**	**Corr.**	**Mean % absolute error**
SentiStrength (standard configuration, 30 runs)	-	60.6%	96.9%	.599	22.0%
Simple logistic regression	700	**58.5%**	96.1%	**.557**	**23.2%**
SVM (SMO)	800	**57.6%**	**95.4%**	**.538**	**24.4%**
J48 classification tree	700	**55.2%**	**95.9%**	.548	24.7%
JRip rule-based classifier	700	**54.3%**	96.4%	**.476**	**28.2%**
SVM regression (SMO)	100	**54.1%**	97.3%	**.469**	**28.2%**
AdaBoost	100	**53.3%**	**97.5%**	**.464**	**28.5%**
Decision table	200	**53.3%**	96.7%	**.431**	**28.2%**
Multilayer Perceptron	100	**50.0%**	*94.1%*	**.422**	**30.2%**
Naïve Bayes	100	**49.1%**	**91.4%**	**.567**	**27.5%**
Baseline	-	**47.3%**	**94.0%**	**-**	**31.2%**
Random	-	**19.8%**	**56.9%**	**.016**	**82.5%**

Bold=sig at 0.01, italic=sig at 0.05 compared to SentiStrength.

For negative sentiment strength, most of the methods give quite similar results and some give better results than SentiStrength. Although the SentiStrength accuracy is 72.8%, this is only 2.9% better than the baseline, several of the other methods have similar levels of accuracy and SVM is significantly more accurate. SentiStrength is significantly the most accurate of the methods if up to one class error is allowed, and has significantly the highest correlation with the human coder results. Note that in theory none of the methods ought to be worse than the baseline but this can occur due to optimisation on the training set rather than the evaluation set. Overall, it seems that SentiStrength is not good at identifying negative emotion but that this is a hard task for the short texts analysed here. Note also that the mean percentage absolute error for the random category is over 100% due to the predominance of '1' as the correct category for negative sentiment.

Table 3. Performance of various algorithms on negative sentiment strength detection for 1,041 comments with the extended feature set and 10-fold cross-validation (decreasing order of positive sentiment strength performance). Other than SentiStrength, results are averages over 4 runs and for the optimal feature numbers, as selected from Figure 2.

Algorithm	**Optimal features**	**Accuracy**	**Accuracy +/- 1 class**	**Corr.**	**Mean % absolute error**
SVM (SMO)	100	*73.5%*	**92.7%**	**.421**	**16.5%**
SVM regression (SMO)	300	73.2%	**91.9%**	**.363**	17.6%
Simple logistic regression	800	72.9%	**92.2%**	**.364**	17.3%
SentiStrength (standard configuration, 30 runs)	-	72.8%	95.1%	.564	18.3%
Decision table	100	72.7%	**92.1%**	**.346**	**17.0%**
JRip rule-based classifier	500	72.2%	**91.5%**	**.309**	17.3%
J48 classification tree	400	71.1%	**91.6%**	**.235**	18.8%
Multilayer Perceptron	100	70.1%	**92.5%**	**.346**	20.0%
AdaBoost	100	**69.9%**	**90.6%**	**-**	**16.8%**
Baseline	-	**69.9%**	**90.6%**	**-**	**16.8%**
Naïve Bayes	200	**68.0%**	**89.8%**	**.311**	**27.3%**
Random	-	**20.5%**	**46.0%**	**.010**	**157.7%**

Bold=sig at 0.01, italic=sig at 0.05 compared to SentiStrength.

The remainder of the paper focuses on positive sentiment alone, since the results for negative sentiment are not significant.

Comparison of feature sets for machine learning –positive sentiment strength

Figures 3 and 4 compare the impact of using different feature sets with the two best-performing algorithms for positive sentiment strength detection. The feature sets are: 1-3-grams; 1-3-grams with emoticons; 1-3-grams with punctuation; 1-3-grams with misspellings (i.e., including terms before spelling correction in addition to terms after spelling correction, when different); 1-3-grams with emoticons, punctuation and misspellings; 1-3-grams with emotion terms; and 1-grams. The basic bag or words approach (1-grams) performs poorly – always the worst feature set for logistic regression and the worst or amongst the worst few feature sets all the time for SVM. For SVM, the best results are achieved with the basic 1-3-grams enhanced by the emotion terms, although most of the time (i.e., for 500-1000 features) the extended feature set (labelled "all of the above" and the same as used in the results above) performs best, perhaps mainly due to the punctuation component, since this enhancement performs second best for 700-1000 features.

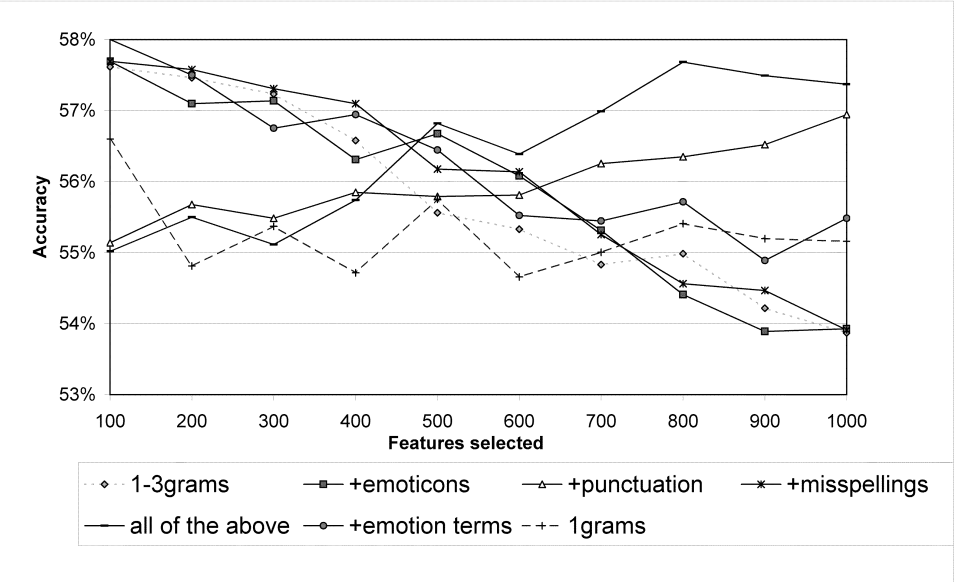

Fig. 3. SVM (SMO) positive sentiment classification accuracy against feature set size for different feature set types; average over 4 classifications.

Figure 4 suggests that, other than the basic bag of words, the difference between feature sets is less clear-cut for logistic regression than for SVM but the best performing combination is again the 1-3 grams plus emotion terms. For larger feature sets, the combined feature set performed best, probably due to the punctuation and emotion terms.

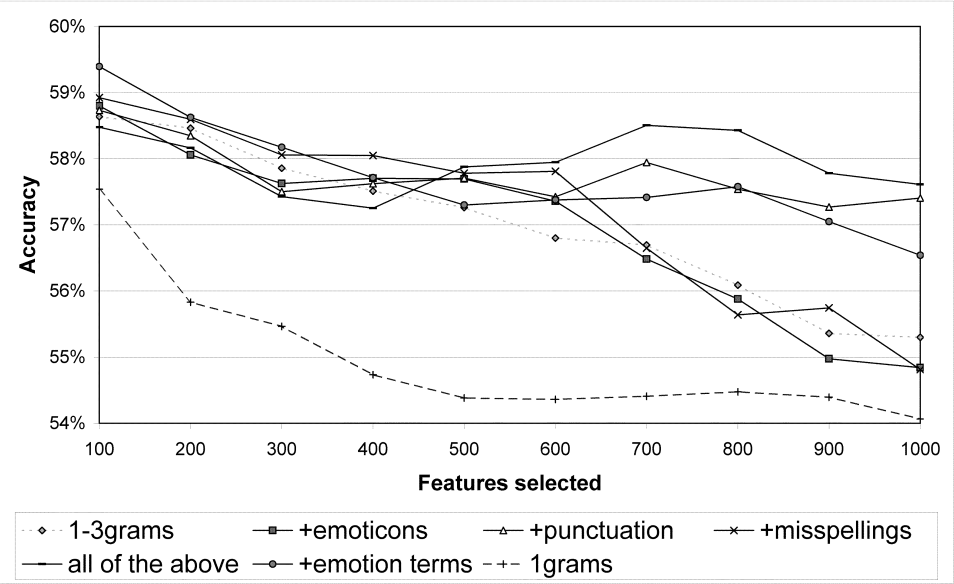

Fig. 4. Logistic regression positive sentiment classification accuracy against feature set size for different feature set types; average over 4 classifications.

A potential weakness of using bigrams and trigrams in conjunction with unigrams is that there is some redundancy involved. For instance, the trigram "I love you" will also match the bigrams "I love" and "love you" as well as the unigrams "I", "love" and "you". In response, subsumption is a feature selection method that eliminates bigrams and trigrams that appear to be redundant in the sense of not giving additional information above that of their constituent unigrams (and bigrams for trigrams). This approach is appropriate here. Subsumption was applied with a logical extension: that word patterns, like happ* could eliminate matching words (e.g., happily, happy in this case) if the appropriate measure was matched. Figures 5 and 6 show the results of subsumption for the two machine learning algorithms for which it performed best: SVM and logistic regression. Subsumption performs best in conjunction with feature reduction, as both graphs show. For the other algorithms, subsumption improved the performance of Jrip by 0.4% ($\alpha = 0.005$, 100 features), SVM regression by 1.1% ($\alpha = 0.02$, 100 features), multilayer perceptron by 1.0% ($\alpha = 0.02$, 100 features) and decision table by 1.0% ($\alpha = 0.005$, 900 features) but did not improve J48, AdaBoost and Naïve Bayes.

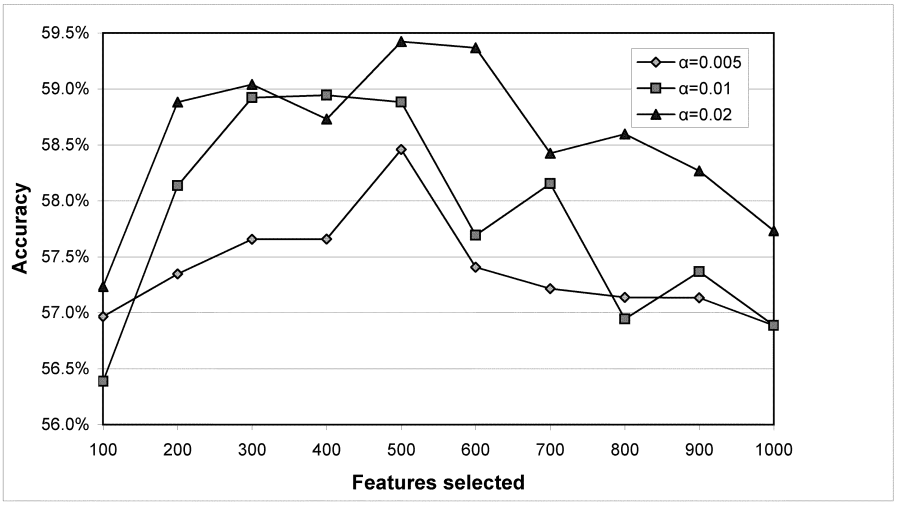

Fig. 5. SVM(SMO) positive sentiment classification accuracy against feature set size for subsumption with various α values; average over 5 classifications.

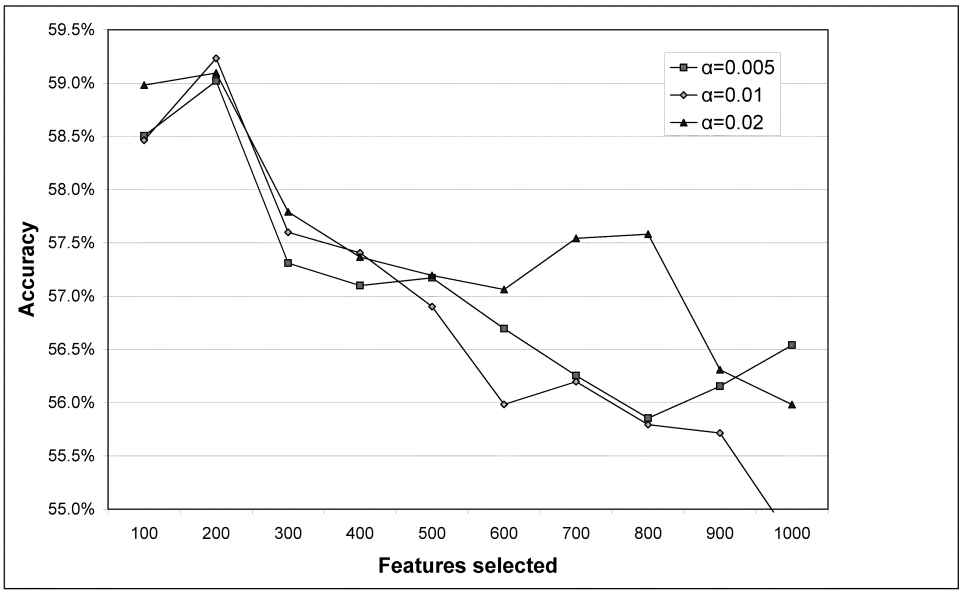

Fig. 6. Logistic regression positive sentiment classification accuracy against feature set size for subsumption with various α values; average over 5 classifications.

From Figure 5, SVM with subsumption outperforms SVM without subsumption on the extended feature set by 1.8%, and outperforms SVM on all the other feature sets (α = 0.02,

500 features). Nevertheless, its accuracy is lower than the SentiStrength standard version, although the difference is not statistically significant (accuracy = 59.42%, accuracy +/-1 = 96.60%, correlation = 0.5822, mean absolute error = 22.65%; only the mean absolute error difference is statistically significant from SentiStrength standard configuration). From Figure 6, logistic regression with subsumption outperforms logistic regression without on the extended feature set by a lower margin of 0.7% ($\alpha = 0.01$, 200 features). It performs less well than 1-3grams with the emotion terms added, however (Figure 4), but this could be a statistical anomaly due to the large number of comparisons performed. Logistic regression performs less well than standard SentiStrength, but the difference is again not significant (accuracy = 59.23%, accuracy +/-1 = 95.79%, correlation = 0.5820, mean absolute error = 22.57%; all except accuracy are statistically significantly different from SentiStrength standard configuration). In terms of α values, 0.02 tends to perform almost uniformly better than other values for this data set.

Note that that although SentiStrength is not statistically significantly better than the optimal SVM and logistic regression models using subsumption, the optimal variation of SentiStrength in Table 4, with one simple modification (training needs only increase of 1 to alter word strengths), is statistically significantly better in all respects than SVM and is statistically significantly better in all respects, except accuracy within +/-1, than logistic regression.

Comparison of SentiStrength versions

Tables 4 and 5 report comparisons of different variations of SentiStrength. Most variations have little influence on the results – individually accounting for a maximum of 0.8% of the performance of the algorithm, except for the last two options. These differences are small enough to be attributable to the corpus used and so the table does not provide convincing evidence that any of the variations are better or worse than the standard approach. When removing all the options (but not changing the averaging method) the cumulative effect is more significant, however, reducing performance by 3.4%. Perhaps comments using non-standard features tend to use multiple non-standard features and so if one special rule is ignored then this is frequently compensated for by the other special rules.

Compared with tables 2 and 3, the main power of SentiStrength is in the combined effect of its rules to adapt to various informal text variations as well as in the overall approach of using a list of term strengths and identifying the strongest positive and negative terms in any comment. In this context, it seems that the generic classification algorithms in Table 2 were a minimum of 2.1% less effective than SentiStrength mainly due to the 1-3 grams approach being insufficiently flexible to cope with non-standard MySpace language (about 3.4% attributable to this cause). In addition, it seems that they were not able to draw upon a large enough training set to learn effective term strengths and a much larger training set could see some of them approach closer to the performance of SentiStrength. Finally, note that the variations of SentiStrength that apparently improve it are not robustly better: when all these are combined to make a new version of SentiStrength this has exactly the same accuracy as the standard configuration (60.64% correct, 97.07% +/- 1 class, .6071 correlation, 21.62% mean % error).

Table 4. Comparison of the **positive** emotion performance over several algorithm variations: average over 30 10-fold cross-validations for 1,041 classified comments.

Type	% Correct	+/- 1 class	corr.	Mean % err. (pred-act)/act
SentiStrength standard algorithm (but training needs only increase of 1 to alter word strengths)	**61.03%**	**96.68%**	.5983	**21.66%**
Negating words *not* used to switch following sentiment (e.g., not happy)	60.87%	**97.50%**	**.6206**	**21.28%**
Multiple consecutive positive words *not* used as emotion boosters	60.70%	96.88%	.5962	21.97%
Emoticons ignored	60.68%	96.87%	.5977	21.95%
Booster words ignored (e.g., very)	60.68%	**96.80%**	.5970	**22.14%**
SentiStrength standard algorithm	60.64%	96.90%	.5986	21.96%
Exclamation marks *not* given a strength of 2	60.51%	**96.62%**	**.6035**	**21.47%**
Automatic spelling correction disabled	*60.39%*	96.88%	*.5961*	22.05%
Extra multiple letters *not* used as emotion boosters	**60.21%**	**96.81%**	**.5952**	**22.16%**
The term "miss" *not* given a strength of +2	60.45%	**96.77%**	**.5953**	**22.16%**
Idiom lookup table disabled	60.52%	96.88%	**.6054**	**21.62%**
Neutral words with emphasis *not* counted as positive emotion	**60.13%**	**96.79%**	.5966	21.90%
SentiStrength with *all* the above changes	**57.44%**	**96.07%**	**.6073**	21.91%
Sentence sentiment is the average of all term sentiments (rather than the maximum)	**42.40%**	**88.54%**	**.4065**	**29.27%**
Text sentiment is the average of all sentence sentiments (rather than the maximum)	**39.13%**	**86.96%**	**.3293**	**33.19%**

* Bold=significant at p=0.01, italic=sig. at p=0.05, compared to the standard algorithm.

Table 5. Comparison of the **negative** emotion performance over several algorithm variations: average over 30 10-fold cross-validations for 1,041 classified comments.

Type	% Correct	+/- 1 class	corr.	Mean % err. (pred-act)/act
Negative sentiment in questions is *not* ignored	**73.56%**	**95.14%**	**.5921**	**18.11%**
SentiStrength standard algorithm (but training needs only increase of 1 to alter word strengths)	72.95%	**94.86%**	.5651	**18.16%**
Negating words *not* used to switch following sentiment (e.g., not happy)	72.84%	**94.79%**	**.5706**	18.35%
SentiStrength standard algorithm	72.83%	95.07%	.5644	18.27%
Multiple consecutive negative words *not* used as emotion boosters	72.81%	95.08%	.5653	18.29%
Emoticons ignored	72.80%	**94.97%**	*.5614*	18.28%
SentiStrength with all the changes in this table except averaging	72.76%	**94.59%**	.5668	**19.07%**
Idiom lookup table disabled	*72.73%*	*95.03%*	**.5556**	**18.63%**
Extra multiple letters *not* used as emotion boosters	72.72%	95.04%	.5627	**18.40%**
Text sentiment is the average of all sentence sentiments (rather than the maximum)	72.66%	**95.83%**	**.5486**	**16.81%**
Automatic spelling correction disabled	**72.64%**	95.07%	**.5586**	**18.62%**
Booster words ignored (e.g., very)	**72.35%**	*95.03%*	**.5559**	**18.50%**
Sentence sentiment is the average of all term sentiments (rather than the maximum)	**72.17%**	**95.35%**	**.4980**	**16.82%**

* Bold=significant at p=0.01, italic=sig.t at p=0.05, compared to the standard algorithm.

Table 5 shows that there is very little variation in the performance of the different variations of SentiStrength for negative emotion strength detection: the performance differs from the standard configuration by a maximum of 0.83%. It suggests however, that negative sentiment in questions (e.g., "Do you hate Tony?") should *not* be ignored in future.

Discussion and Conclusions

Recall that the main novel contributions of this paper are: a machine learning approach to optimise sentiment term weightings; methods for extracting sentiment from non-standard spelling in text; and a related spelling correction method. SentiStrength was able to identify the strength of positive sentiment on a scale of 1 to 5 in 60.6% of the time in informal MySpace language, significantly above the best standard machine-learning approaches which had a performance of up to 58.5% - in line with those for a previous 4-category opinion intensity classification task (Wilson et al., 2006). The standard version of SentiStrength was also better then standard machine learning methods when their performance was improved (or not, in some cases) with the use of subsumption and information gain feature reduction, but the difference was not statistically significant. A slightly modified version of SentiStrength was statistically significantly better than the improved machine learning methods, however. This is good evidence of the efficacy of SentiStrength for positive sentiment strength detection given the range of different algorithms and parameters that it was compared against (9 algorithms x 11 feature set sizes, x 7 feature set types = 693 variations, plus 9 algorithms x 10 feature set sizes x 3 α values = 270 variations for subsumption), which gives lower-performing algorithms a reasonable statistical chance of outperforming SentiStrength through chance, but none did.

The main reason for SentiStrength's relative success seems to be procedures for decoding non-standard spellings and methods for boosting the strength of words, which accounted for much of its performance. Without these factors, the SentiStrength variant based solely upon a dictionary of emotion-associated words and their estimated strengths with 57.5% was only 1.3% better than the most successful machine learning approach on an extended set of 1-3grams. In contrast, SentiStrength was able to identify negative sentiment little better (1.8%) than the baseline, probably due to creativity in expressing negative comments or due to the difficulty in getting significantly above the baseline when one category dominates (Artstein & Poesio, 2008; Krippendorff, 2004). It seems that both positive and negative sentiment detection in informal text language like MySpace comments is challenging because of several factors: language creativity, expressions of sentiment without emotion-bearing words, and differences between human coder interpretations meaning that there is not a genuinely correct classification for most comments.

Given the success in generating an algorithm for positive sentiment strength detection and the predominance of positive sentiment in MySpace comments, it seems that future research can apply the sentiment strength detection techniques to automatically identify and classify positive sentiment in informal web communication environments on a large scale. Moreover, there are many commercial applications of sentiment analysis, some of which use informal computer text generate from chatrooms or mobile phone text messages, and this algorithm shows that it is possible to estimate the strength of positive sentiment even in these short messages.

In terms of future work, a next logical step is to attempt to improve the performance of the system through linguistic processing, despite the poor grammar of the short informal text messages analysed. Previous work has shown that this approach is promising, particularly via dependency trees (Wilson et al., 2009) and that, given a large enough training sample, improvements may be possible even in poor quality text (Gamon, 2004).

Appendix: Coder Instructions (extract)

Code each comment for the degree to which it expresses positive emotion *or energy*. Excitement, enthusiasm or energy should be counted as positive emotion here. If you think

that the punctuation emphasises the positive emotion or energy in any way then include this in your rating. The scale for ***positive*** emotion or energy is:

[no positive emotion or energy] **1– 2 – 3 – 4 – 5** [very strong positive emotion]

- Allocate 1 if the comment contains no positive emotion or energy.
- Allocate 5 if the comment contains very strong positive emotion.
- Allocate a number between 2 and 4 if the comment contains some positive emotion but not very strong positive emotion. Use your judgement about the exact positive emotion strength.

Code each comment for the degree to which it expresses negative emotion or is negative. If you think that the punctuation emphasises the negative emotion in any way then include this in your rating. The scale for ***negative*** emotion is:

[no negative emotion] **1– 2 – 3 – 4 – 5** [very strong negative emotion]

- Allocate 1 if the comment contains no negative emotion at all.
- Allocate 5 if the comment contains very strong negative emotion.
- Allocate a number between 2 and 4 if the comment contains some negative emotion but not very strong negative emotion. Use your judgement about the exact negative emotion strength.

When making judgements, please be as consistent with your previous decisions as possible. Also, please interpret emotion within the individual comment that it appears and ignore all other comments.

References

Abbasi, A., Chen, H., & Salem, A. (2008). Sentiment analysis in multiple languages: Feature selection for opinion classification in web forums. *ACM Transactions on Information Systems, 26*(3), 12:11-12.34.

Abbasi, A., Chen, H., Thoms, S., & Fu, T. (2008). Affect analysis of Web forums and Blogs using correlation ensembles. *IEEE Transactions on Knowledge and Data Engineering, 20*(9), 1168-1180.

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.

Argamon, S., Whitelaw, C., Chase, P., Hota, S. R., Garg, N., & Levitan, S. (2007). Stylistic text classification using functional lexical features. *Journal of the American Society for Information Science and Technology, 58*(6), 802-822.

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

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.

Baron, N. S. (2003). Language of the Internet. In A. Farghali (Ed.), *The Stanford Handbook for Language Engineers* (pp. 59-127). Stanford: CSLI Publications.

Barrett, L. F. (2006). Valence as a basic building block of emotional life. *Journal of Research in Personality, 40*(1), 35-55.

boyd, d. (2008). *Taken out of context: American teen sociality in networked publics.* University of California, Berkeley, Berkeley.

boyd, d. (2008). Why youth (heart) social network sites: The role of networked publics in teenage social life. In D. Buckingham (Ed.), *Youth, identity, and digital media* (pp. 119-142). Cambridge, MA: MIT Press.

Brill, E. (1992). A simple rule-based part of speech tagger. *Proceedings of the Third Conference on Applied Natural Language Processing*, 152-155.

Chaumartin, F.-R. (2007). UPAR7: A knowledge-based system for headline sentiment tagging. In *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007)* (pp. 422-425). New York, NY: ACM.

Choi, Y., & Cardie, C. (2008). Learning with compositional semantics as structural inference for subsentential sentiment analysis. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 793-801.

Cohen, W. (1995). Fast effective rule induction. *Proceedings of the Twelfth International Conference on Machine Learning*, 115–123.

Cornelius, R. R. (1996). *The science of emotion*.Upper Saddle River, NJ: Prentice Hall.

Crystal, D. (2006). *Language and the Internet* (2nd ed.). Cambridge, UK: Cambridge University Press.

Das, S., & Chen, M. (2001). Yahoo! for Amazon: Extracting market sentiment from stock message boards. *Proceedings of the Asia Pacific Finance Association Annual Conference (APFA)*, Bangkok, Thailand, July 22-25, Retrieved July 17, 2009 from: http://sentiment.technicalanalysis.org.uk/DaCh.pdf.

Denecke, K., & Nejdl, W. (2009). How valuable is medical social media data? Content analysis of the medical web. *Information Sciences, 179*(12), 1870-1880.

Derks, D., Bos, A. E. R., & von Grumbkow, J. (2008). Emoticons and online message interpretation. *Social Science Computer Review, 26*(3), 379-388.

Derks, D., Fischer, A. H., & Bos, A. E. R. (2008). The role of emotion in computer-mediated communication: A review. *Computers in Human Behavior, 24*(3), 766–785.

Diener, E., & Emmons, R. A. (1984). The independence of positive and negative affect. *Journal of Personality and Social Psychology, 47*(5), 1105-1117.

Ekman, P. (1992). An argument for basic emotions. *Cognition and Emotion, 6*(3/4), 169-200.

Esuli, A., & Sebastiani, F. (2006). SENTIWORDNET: A publicly available lexical resource for opinion mining. *Proceedings of Language Resources and Evaluation (LREC) 2006*, Retrieved July 28, 2009 from: http://tcc.fbk.eu/projects/ontotext/Publications/LREC2006-esuli-sebastiani.pdf.

Fox, E. (2008). *Emotion science*.Basingstoke: Palgrave Macmillan.

Fullwood, C., & Martino, O. I. (2007). Emoticons and impression formation. *The Visual in Popular Culture, 19*(7), 4-14.

Gamon, M. (2004). Sentiment classification on customer feedback data: noisy data, large feature vectors, and the role of linguistic analysis. *Proceedings of the 20th international conference on Computational Linguistics*, No.841.

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

Gill, A. J., Gergle, D., French, R. M., & Oberlander, J. (2008). Emotion rating from short blog texts. In *Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems* (pp. 1121-1124). New York, NY: ACM.

Grinter, R. E., & Eldridge, M. (2003). Wan2tlk? everyday text messaging. *CHI 2003,* 441-448.

Hancock, J. T., Gee, K., Ciaccio, K., & Lin, J. M.-H. (2008). I'm sad you're sad: Emotional contagion in CMC. *Proceedings of the ACM 2008 conference on Computer supported cooperative work*, 295-298.

Hopkins, D. J., & King, G. (2010). A method of automated nonparametric content analysis for social science. *American Journal of Political Science, 54*(1), 229-247.

Huang, Y.-P., Goh, T., & Liew, C. L. (2007). Hunting suicide notes in web 2.0 - Preliminary findings. In *Ninth Ieee International Symposium On Multimedia - Workshops, Proceedings* (pp. 517-521). Los Alamitos: IEEE.

Huppert, F. A., & Whittington, J. E. (2003). Evidence for the independence of positive and negative well-being: Implications for quality of life assessment. *British Journal of Health Psychology, 8*(1), 107-122.

Kaji, N., & Kitsuregawa, M. (2007). Building lexicon for sentiment analysis from massive collection of HTML documents. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning* (pp. 1075-1083, retrieved July 1028 from: http://www.aclweb.org/anthology/D/D1007/D1007-1115.pdf).

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

Kukich, K. (1992). Techniques for automatically correcting words in text. *ACM computing surveys, 24*(4), 377-439.

Liu, H., Lieberman, H., & Selker, T. (2003). A model of textual affect sensing using real-world knowledge. *Proceedings of the 2003 International Conference on Intelligent User Interfaces, IUI 2003*, 125-132.

Mauss, I. B., & Robinson, M. D. (2009). Measures of emotion: A review. *Cognition and Emotion, 23*(2), 209-237.

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*.

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.

Nardi, B. A. (2005). Beyond bandwidth: Dimensions of connection in interpersonal communication. *Computer-Supported Cooperative Work, 14*(1), 91-130.

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

Ng, V., Dasgupta, S., & Arifin, S. M. N. (2006). Examining the role of linguistic knowledge sources in the automatic identification and classification of reviews. *Proceedings of the COLING/ACL 2006 Main Conference*, 611-618.

Pang, B., & Lee, L. (2004). Sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of ACL 2004* (pp. 271-278). New York: ACL Press.

Pang, B., & Lee, L. (2005). Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. *Proceedings of the 43rd Annual Meeting of the ACL*, 115-124.

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.

Pennebaker, J. W., Mayne, T., & Francis, M. E. (1997). Linguistic predictors of adaptive bereavement. *Journal of Personality and Social Psychology, 72*(4), 863-871.

Pollock, J. J., & Zamora, A. (1984). Automatic spelling correction in scientific and scholarly text. *Communications of the ACM, 27*(4), 358-368.

Prabowo, R., & Thelwall, M. (2009). Sentiment analysis: A combined approach. *Journal of Informetrics, 3*(1), 143-157.

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.

Riloff, E., & Wiebe, J. (2003). Learning extraction patterns for subjective expressions. *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing (EMNLP-03)*, Retrieved April 11, 2010 from: http://www.cs.utah.edu/~riloff/pdfs/emnlp2003.pdf.

Russell, J. A. (1979). Affective space is bipolar. *Journal of Personality and Social Psychology, 37*(3), 345-356.

Schapire, R., & Singer, Y. (2000). BoosTexter: A boosting-based system for text categorization. *Machine Learning, 39*(2/3), 135-168.

Short, J. C., & Palmer, T. B. (2008). The application of DICTION to content analysis research in strategic management. *Organizational Research Methods, 11*(4), 727-752.

Snyder, B., & Barzilay, R. (2007). Multiple aspect ranking using the good grief algorithm. *Proceedings of NAACL HLT*.

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.

Stoppard, J. M., & Gunn Gruchy, C. D. (1993). Gender, context, and expression of positive emotion. *Personality and Social Psychology Bulletin, 19*(2), 143-150.

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.

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. (2009). MySpace comments. *Online Information Review, 33*(1), 58-76.

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.

Thurlow, C. (2003). Generation Txt? The sociolinguistics of young people's text-messaging. *Discourse Analysis Online, 1*(1), Retrieved January 3, 2008 from: http://extra.shu.ac.uk/daol/articles/v2001/n2001/a2003/thurlow2002003-paper.html.

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.

Walther, J., & Parks, M. (2002). Cues filtered out, cues filtered in: computer-mediated communication and relationships. In M. Knapp, J. Daly & G. Miller (Eds.), *The Handbook of Interpersonal Communication (3rd ed.)* (pp. 529-563). Thousand Oaks, CA: Sage.

Watson, D. (1988). Intraindividual and interindividual analyses of positive and negative affect: their relation to health complaints, perceived stress, and daily activities. *Journal of Personality and Social Psychology, 54*(6), 1020-1030.

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology, 54*(6), 1063-1070.

Wiebe, J., Wilson, T., Bruce, R., Bell, M., & Martin, M. (2004). Learning subjective language. *Computational Linguistics, 30*(3), 277-308.

Wiebe, J., Wilson, T., & Cardie, C. (2005). Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation, 39*(2-3), 165-210.

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

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, C.-H., Chuang, Z.-J., & Lin, Y.-C. (2006). Emotion recognition from text using semantic labels and separable mixture models. *ACM Transactions on Asian Language Information Processing, 5*(2), 165-183.