Damping Sentiment Analysis in Online Communication: Discussions, Monologs and Dialogs

Mike Thelwall[1], Kevan Buckley[1], George Paltoglou[1], Marcin Skowron[2], David Garcia[3], Stephane Gobron[4], Junghyun Ahn[4], Arvid Kappas[5], Dennis Küster[5], Janusz A. Holyst[6]

[1]Statistical Cybermetrics Research Group, University of Wolverhampton, Wolverhampton, UK
{m.thelwall, K.A.Buckley, g.paltoglou}@wlv.ac.uk
[2]Austrian Research Institute for Artificial Intelligence, Vienna, Austria
marcin.skowron@ofai.at
[3]Chair of Systems Design, ETH Zurich, Weinbergstrasse 56/58, 8092 Zurich, Switzerland
dgarcia@ethz.ch
[4]Information and Communication Systems Institute (ISIC), HE-Arc, HES-SO, Switzerland
stephane.gobron@gmail.com
[4]SCI IC RB Group, Ecole polytechnique fédérale de Lausanne EPFL, Switzerland
junghyun.ahn@epfl.ch
[5]School of Humanities and Social Sciences, Jacobs University Bremen, Bremen, Germany
{a.kappas, d.kuester}@jacobs-university.de
[6]Center of Excellence for Complex Systems Research, Faculty of Physics, Warsaw University of Technology, Warsaw, Poland
jholyst@if.pw.edu.pl

Abstract. Sentiment analysis programs are now sometimes used to detect patterns of sentiment use over time in online communication and to help automated systems interact better with users. Nevertheless, it seems that no previous published study has assessed whether the position of individual texts within on-going communication can be exploited to help detect their sentiments. This article assesses apparent sentiment anomalies in on-going communication – texts assigned significantly different sentiment strength to the average of previous texts – to see whether their classification can be improved. The results suggest that a damping procedure to reduce sudden large changes in sentiment can improve classification accuracy but that the optimal procedure will depend on the type of texts processed.

Keywords: Sentiment analysis, opinion mining, social web.

1 Introduction

The rapid development of sentiment analysis in the past decade has roots in the widespread availability of social web texts that are relevant to marketing needs. In particular, formal or informal product reviews online can now be mined with a wide range of sentiment analysis programs in multiple languages to give businesses information about what the public thinks about products and brands (Liu, 2012; Pang & Lee, 2008). By harnessing real-time sources like Twitter, businesses can even be given daily updates about changes in average sentiment. More recently, however, sentiment analysis programs have been used to identify the sentiment expressed in texts, irrespective of whether any products are mentioned. One goal of this type of research has been to identify trends in sentiment over time in relation to a specific topic (Chmiel et al., 2011a; Garas, Garcia, Skowron, & Schweitzer, 2012) or more generally (Thelwall, Buckley, & Paltoglou, 2011) or in a particular genre (Dodds & Danforth, 2010; Kramer, 2010): both social sciences types of research. Another type of research detects users' sentiments in order to react to them in real time. As an example of the latter, dialog systems have been developed that react differently to users depending on the sentiment expressed (Skowron, 2010) and in one online environment, the facial expressions of an automatic chat partner in a three-dimensional virtual world respond to the sentiment expressed by the participants, as detected with a sentiment analysis program (Gobron et al., 2011; Skowron et al., 2011). In another computing application that is somewhat similar to this, the Yahoo! Answers system harnesses sentiment analysis to help identify people that receive positive feedback after submitting their answers so that these people can be identified and their answers given prominence in search results (Kucuktunc, Cambazoglu, Weber, & Ferhatosmanoglu, 2012). As a result of such applications, there is a need for sentiment analysis software that is optimised for general social web texts and that can take advantage of any regular patterns of sentiment expressions and reactions online in order to improve the accuracy of the predictions made.

Some research from psychology and from studies of online communication can shed light on how sentiment is best detected and measured in online environments. Psychologists have investigated emotions for over a century and today there is a field of emotion psychology (Cornelius, 1996; Fox, 2008). One important finding is that humans seem to process positive and negative sentiment separately and relatively independently. This means that although it is often practical and convenient to measure positive and negative sentiment together to give one combined overall result for each text, it is more natural to measure them separately and report two scores per text. Psychology research also confirms that emotions vary in strength (Cornelius, 1996; Fox, 2008) and so the natural way to measure emotion and hence sentiment is on a dual scale measuring the strength of positive and negative sentiment expressed. Emotion psychologists also recognize a range of different types of emotion (e.g., anger, hate) rather than just positivity and negativity but studies suggest that the fundamental divide is between positive and negative emotion with more fine-grained emotions being socially constructed to some extent (Fox, 2008). Thus it is reasonable from a psychology perspective to either focus on positive and negative sentiment or on more fine-grained sentiment, with the latter probably reflecting social conditioning more.

Research from non-psychologists has investigated emotion and sentiment online to see whether there are patterns in the use of sentiment in ongoing communications, with positive results. A common finding is that whilst different social web environment have different average levels of positive and negative sentiment (e.g., political discussions tend to be negative whereas comments between friends tend to be positive) (Thelwall, Buckley, & Paltoglou, 2012) above average levels of negativity associate with longer interactions: negativity seems to fuel longer discussions (Chmiel et al., 2011ab; Thelwall, Sud, & Vis, 2012). Additionally, and perhaps unsurprisingly, some studies have found evidence of sentiment homophily between online friends: people tend to express similar levels of sentiment to that expressed by their friends, compared to the overall average (Bollen, Pepe, & Mao, 2011; Thelwall, 2010).

The above discussion suggests that the task of sentiment analysis in general social web texts may need to be tackled somewhat differently to that of product review sentiment analysis or opinion mining. Whilst there are programs, such as SentiStrength (discussed below), that are designed for social web texts it seems that all process each text separately and independently and none have attempted to improve sentiment detection by taking advantage of patterns of online communication, although some have successfully exploited discourse features (Somasundaran, Namata, Wiebe, & Getoor, 2009). This article assesses the potential for improving sentiment detection in this way. As an exploratory study, it uses four different types of social web context for evaluations (political forum discussions, non-political forum discussions, as well as dialogs and monologs in Twitter). It also assesses one simple method of exploiting the sentiment of previous texts when classifying the sentiment of new texts: damping. Defined precisely below, the damping method changes a sentiment prediction by bringing it closer to the average sentiment of the previous few texts if the prediction would otherwise be too different from this average. The experimental results suggest that the damping method works well in some contexts but not all and so should be used with care.

2 Sentiment analysis

Previous sentiment analysis or opinion mining research has used many different methods in order to detect the sentiment of a text or the opinion expressed in a text towards a product or an aspect of a product. Lexical methods typically start with a pre-defined lexicon of terms with known typical sentiment polarity, such as SentiWordNet (Baccianella, Esuli, & Sebastiani, 2010), sentiment terms from the General Inquirer lexicon (Choi & Cardie, 2008), LIWC (Pennebaker, Mehl, & Niederhoffer, 2003) as in (Thelwall, Buckley, Paltoglou, Cai, & Kappas, 2010), or a human-created list of sentiment terms (Taboada, Brooke, Tofiloski, Voll, & Stede, 2011). These lists are then matched with terms in texts to be classified and then a set of rules applied to classify the texts. Classifications are typically either binary (positive or negative), or trinary (positive, negative or neutral/objective) although some also detect sentiment strength in addition to polarity.

A non-lexical approach is to use machine learning methods to decide which words are the most relevant for sentiment based upon a set of linguistic or non-linguistic rules and a large set of pre-classified texts for training. An advantage of not using a pre-defined lexicon, which is particularly relevant when developing a sentiment classifier for reviews of a particular type of product, is that non-sentiment terms may be identified that carry implied sentiment by expressing a judgment, such as "heavy" in the phrase "the phone was very heavy". The limitation of needing a corpus of human-coded texts to train a non-lexical classifier can be avoided in some cases by exploiting free online product review sites in which reviewers score products in addition to giving text reviews. In the absence of these, other unsupervised methods (Turney, 2002) and domain transfer methods (Glorot, Bordes, & Bengio, 2011; Ponomareva & Thelwall, 2012) have also been developed. Two disadvantages of the non-lexical approach for social science research purposes, however, are that they can introduce systematic anomalies through exploiting non-sentiment words (Thelwall et al., 2010) and that they seem to be less transparent than lexical methods, which can often give a clear explanation as to why a sentence has been classified in a certain way, by reference to the predefined list of sentiment terms (e.g., "this sentence was classified as positive because it contains the word 'happy', which is in the lexicon of positive terms"). Sentiment analysis methods can exploit linguistic structure to make choices about the types of words to analyze, such as just the adjectives (Wiebe, Wilson, Bruce, Bell, & Martin, 2004).

Although most sentiment analysis programs seem to classify entire texts as positive, negative or neutral, aspect-based sentiment analysis classifies texts differently based upon the aspects of a product discussed. For instance, an aspect-based classifier might detect that "cheap" is negative in the context of a phone design but positive in the context of the phone's price. Other programs are more fine-grained in a different sense: classifying multiple emotions, such as anger, sadness, hate, joy and happiness (Neviarouskaya, Prendinger, & Ishizuka, 2010) and/or sentiment strength (Wilson, Wiebe, & Hwa, 2006).

Some sentiment analysis programs have attempted to use the position of a text in order to help classify sentiment, but only for the larger texts containing classified smaller texts. In movie reviews, sentences near the end typically carry more weight than earlier sentences and hence movie review classifiers that work by detecting the sentiment of individual sentences and then aggregating the results to predict the sentiment of the overall review can improve their performance by giving higher weights to later texts (Pang, Lee, & Vaithyanathan, 2002). Discourse structure has been successfully used in one case to classify contributions in work-based meetings as positive, negative or neutral, producing a substantial increase in accuracy in comparison to baseline approaches (Somasundaran et al., 2009). This promising approach has not been tried for social web texts, however, and may work best in formal discussions. Another investigation uses discourse structure to help separate discussion participants into different camps but not to help classify the sentiment of their texts (Agrawal, Rajagopalan, Srikant, & Xu, 2003). Despite these examples, no sentiment analysis seem to exploit the occurrence of many texts in communication chains, such as monologs, dialogs or multi-participant discussions, in order to predict their sentiment more accurately.

3 Sentiment strength detection with SentiStrength

The damping method described below was tested by being applied to SentiStrength (Thelwall & Buckley, in press; Thelwall et al., 2010; Thelwall et al., 2012). This sentiment analysis program was chosen because it is designed to detect the strength of positive and negative sentiment in short informal text and has been tested on a range of different social web text types: Tweets, MySpace comments, RunnersWorld forum posts, BBC discussion forum posts, Digg posts, and comments on YouTube videos. SentiStrength assigns a score of 1, 2, 3, 4, or 5 for the strength of positive sentiment and -1, -2, -3, -4 or -5 for the strength of negative sentiment, with each text receiving one score for each. For instance, the text "I hate Tony but like Satnam" might get a score of (-4, 3), indicating strong negative sentiment and moderate positive sentiment.

SentiStrength's dual positive/negative scoring scheme is unusual for sentiment strength detection and stems from the psychology input to the design of the software because psychologists accept that humans process positive and negative sentiment in parallel rather than in a combined way (Norman et al., 2011); hence positive and negative sentiment do not necessarily cancel each other out. As mentioned above, for a psychological analysis of sentiment, and hence for a social science analysis of sentiment, it is reasonable to detect positive and negative sentiment separately. SentiStrength has been used to analyze social web texts to detect patterns of communication but no previous study has attempted to improve its performance by taking advantage of sentiment patterns in on-going communications.

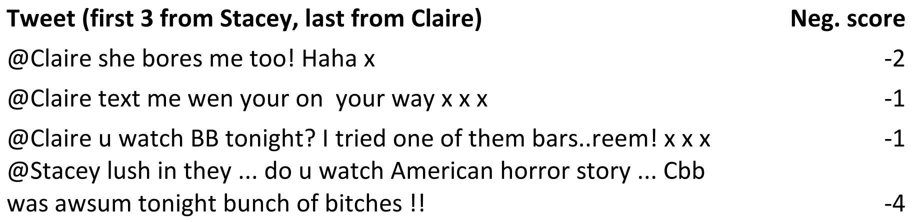
SentiStrength works primarily through a lexicon of terms with positive and negative weights assigned to them. In the above example, "hate" is in the lexicon with strength -4 and "like" has strength +3. Each text is given a score equal to the largest positive and negative value of the sentiment words contained in it, subject to some additional rules. These rules include methods for dealing with negation (e.g., don't), booster words (e.g., very), emoticons, and informal expressions of sentiment (e.g., "I'm haaaaaapy!!!").

3.1 Sentiment damping

The adjustment method is based upon the assumption that a text in a series that has a significantly different sentiment level than the previous texts, according to a classifier, may be an anomaly in the sense of having been misclassified and may have a real sentiment that is closer to the average. This is operationalized by two rules:

- If the classified positive sentiment of text A differs by at least 1.5 from the average positive sentiment of the previous 3 posts, then adjust the positive sentiment prediction of text A by 1 point to bring it closer to the positive average of the previous 3 terms.
- If the classified negative sentiment of text A differs by at least 1.5 from the average negative sentiment of the previous 3 posts, then adjust the negative sentiment prediction of text A by 1 point to bring it closer to the negative average of the previous 3 terms.

For example, if four consecutive texts are classified as 1, 2, 1, 4 for positive sentiment then rule 1 would be triggered since 4 is more than 2 greater than the average of 1, 2, and 1, and hence the prediction of 4 would be adjusted by 1 towards the average. Hence the adjusted predictions would be 1, 2, 1, 3. Figure 1 is another example from the Twitter dialogs data set.



Tweet (first 3 from Stacey, last from Claire)	Neg. score
@Claire she bores me too! Haha x	-2
@Claire text me wen your on your way x x x	-1
@Claire u watch BB tonight? I tried one of them bars..reem! x x x	-1
@Stacey lush in they ... do u watch American horror story ... Cbb was awsum tonight bunch of bitches !!	-4

Fig. 1. A dialog between two tweeters with SentiStrength negative classifications that would trigger damping for the final contribution. The term *horror* triggered a strong negative score in the final contribution but human coders judged that this was not strongly negative, presumably because it was part of a TV series name. This type of anomaly would be corrected by the damping method (names changed and contributions slightly changed to anonymize participants).

4 Data sets

Multiple data sets were created to reflect different kinds of web-based informal communication: discussions, dialogs and monologs.

4.1 BBC World news discussions (BWNpf)

This data set consists of contributions to the BBC World News online discussion forum. This was chosen as an example of a political forum discussion in which multiple participants can contribute. Contributions were selected for coding if the adjustment rule would trigger a positive or negative change in them. In addition, a random set of non-adjusted texts was also selected for coding. A text was not chosen if any of the previous 3 contributions to the discussion had been chosen. This was to avoid taking too many contributions from the same part of the discussion.

4.2 RunnersWorld (RWtf)

This data set consists of contributions to the RunnersWorld online marathon running discussion forum. This was chosen as an example of a non-political topical discussion forum in which multiple participants can contribute. Although the forum focuses on a single topic, this is probably true for most online discussion forums and so it represents a popular type of online discussion despite its specialist nature. Contributions were selected in the same way as for the BWNd data set.

4.3 Twitter monologs (Tm)

This data set consists of tweets in English from randomly selected Twitter users tweeting in English and geolocated in the US. This data set was obtained by monitoring the Twitter API with a blank US geolocation search during early 2012. Each "monolog" in the dataset consists of all tweets from the random user, and at least 10 tweets per user. This represents tweeting in the sense of broadcasting comments rather than necessarily interacting with other tweeters, although some comments may also be interactions. Tweets were selected for coding as for BWNd.

4.4 Twitter dialogs (Td)

This data set is similar to Td but represents a set of dialogs between pairs of users. For each user in the Tm data set, a random target (i.e., a Tweeter, indicated using the @ convention) of one of their tweets was selected and all of this user's tweets were downloaded. If the target user also targeted the original user then their tweets were combined and arranged in chronological order to form a Twitter "dialog" in this data set, discarding all tweets not directed at the other dialog partner. For instance, if the two contributors were User1 and User2, then tweets from User1 were discarded unless they contained @User2 and tweets from User2 were discarded unless they contained @User1. Contributions were randomly selected from these dialogs for coding subject to the restriction that a contribution must be either preceded to followed by a contribution from the other dialog participant (so that they would not be part of a mini-monolog rather than a genuine dialog).

4.5 Preliminary analysis of data sets

Table 1 reports some basic statistics from SentiStrength (without damping) applied to the four data sets. The table reports the average of all statistics calculated separately for each thread/monolog/dialog in each sample. The results show differences between the data sets in all statistics. For example, the RunnersWorld forum threads have the highest average positive sentiment strength and the BBC World News forum has the highest average negative sentiment strength, probably reflecting their discussion topics. The negative correlations between positive and negative scores for the first two data sets in comparison to positive correlations between positive and negative scores last two probably reflects the length limit on tweets: a slight tendency for tweets to contain either positive or negative sentiment but not both. In contrast, for the first two forums, if a person expresses negative sentiment then they are also likely to express positive sentiment and vice versa. This would be consistent with some texts being factual or objective and others being subjective.

Of most interest here are the lag 1 autocorrelations: these are correlations between the sentiment scores and the sentiment scores offset by one. High correlations (close to 1) would suggest that the sentiment of a post tends to be similar to the sentiment of the previous post, supporting the damping method for sentiment analysis. Although all the autocorrelations are significantly non-zero they seem to be small enough to be irrelevant in practice. This suggests that within these data sets, texts with similar sentiment levels have only a small tendency to cluster together.

Table 1. Statistics and autocorrelations for the threads/monologs/dialogs with at least 30 contributions. All correlations and autocorrelations are significantly different from 0 at p=0.001

Data set	Sample size*	Mean positive	Mean negative	Positive-negative correlation	Lag 1 positive autocorr.	Lag 1 negative autocorr.
BWNpf	4580	1.918	-2.414	-.2378	.0331	.0529
RWtf	4958	2.200	-1.666	-.1867	.0924	.0634
Tm	675	1.691	-1.364	.0328	.0558	.0529
Td	329	1.778	-1.367	.0349	.0299	.0389

* Sample size is number of threads for BWNpf and RWtf, the number of dialogs for Tm and the number of monologs for Td.

4.6 Inter-coder consistency

The texts selected as described above for each data set were given to two experienced coders who were not associated with the project and who were not told the purpose of the project. The coders were given the texts to code, along with the previous texts in the dialog/monolog/thread in order to reveal the context of each text for more accurate coding. The coders were asked to score each text with the standard SentiStrength scheme of two whole numbers: [no positive sentiment] 1 – 2 – 3 – 4 – 5 [very strong positive sentiment] and [no negative sentiment] -1 – -2 – -3 – -4 – -5 [very strong negative sentiment]. The coders were each given a standard codebook to describe and motivate the task and were requested to code for a maximum of one hour per day, to minimise the risk of mistakes through fatigue.

Krippendorff's inter-coder weighted alpha (Krippendorff, 2004) was used to calculate the extent of agreement between the coders, using the difference between the categories assigned as the weights. The results showed that the level of inter-coder agreement was good but not excellent, probably because sentiment is a subjective phenomenon. It is therefore reasonable to use the values of the coders to assess the sentiment analysis results. The values of the second coder were chosen because this person coded more texts.

Table 2. Krippendorff inter-coder weighted alpha values for the similarity between codes from the two coders.

Data set	Positive sentiment α	Negative sentiment α
BWNpf (n=466)	0.655	0.559
RWtf (n=379)	0.572	0.659
Tm (n=445)	0.695	0.744
Td (n=508)	0.689	0.738

5 Experimental Results

Table 3 reports a comparison of the results for damped SentiStrength with undamped SentiStrength for the random selection of human coded texts that were damped by SentiStrength (i.e., only the changed values). The table reports damping increases in sentiment strength separately from damping decreases in sentiment strength. For each type of damping, the result is either a more accurate or a less accurate prediction and Table 3 reports the proportion of each. The results are mixed: an overall improvement in 9 of the 16 cases examined (although three are marginal: 51%, 51% and 54%) and no clear pattern about which of the four types of damping are always effective. Nevertheless, there are six cases in which the improvement is substantial – 65% to 75% – and this suggests that if damping is applied selectively by choosing which of the four types to use for a given data set then this should improve sentiment classification accuracy.

Table 3. Percentage of sentiment classification improvements when damping increases sentiment scores and when damping decreases sentiment scores. Figures above 50% indicate an overall increase in classification accuracy.

Data set	Positive sentiment increase improvement	Positive sentiment decrease improvement	Negative sentiment increase improvement	Negative sentiment decrease improvement
BWNpf	38% (n=74)	73% (n=127)	75% (n=165)	51% (n=166)
RWtf	71% (n=175)	43% (n=153)	54% (n=139)	65% (n=280)
Tm 5b	71% (n=97)	33% (n=319)	51% (n=55)	41% (n=300)
Td	69% (n=81)	33% (n=304)	47% (n=43)	44% (n=331)

6 Conclusions

The results clearly show that damping can improve sentiment strength detection for social web texts, although some forms of damping have no effect on particular types of text or make the results worse. Hence, when optimising sentiment analysis for a new dataset, experiments should be run to decide which of the four types of damping to include and which to exclude (i.e., damping sentiment increases, damping sentiment decreases, for both positive and negative sentiment). A limitation of this approach is that the performance improvement caused by damping is likely to be minor because only a minority of predictions will be damped, depending on the corpus used. Moreover, a practical limitation is that human-coded texts will be needed to identify the types of damping to use. This human coding is resource-intensive because it must be conducted specifically for the damping, with a dataset of texts potentially subject to damping changes, and hence would not be a random set of texts that could be used for other evaluations.

For future work, it would be useful to conduct a larger scale and more systematic evaluation of different types of texts in order to produce recommendations for the contexts in which the different types of damping should be used. This would save future researchers the time needed to test each new data set to select which damping methods to use. It would also be useful to compare this approach to the use of discourse markers (Somasundaran et al., 2009) and attempt to combine both to improve on the performance of each one.

7 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).

8 References

1. Agrawal, R., Rajagopalan, S., Srikant, R., & Xu, Y. (2003). Mining newsgroups using networks arising from social behavior. *Proceedings of WWW,* 529-535.
2. 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/769_Paper.pdf.
3. 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.
4. Chmiel, A., Sienkiewicz, J., Thelwall, M., Paltoglou, G., Buckley, K., Kappas, A., & Hołyst, J. A. (2011a). Collective emotions online and their influence on community life. *PLoS ONE, 6*(7), e22207.
5. Chmiel, A., Sienkiewicz, J., Paltoglou, G., Buckley, K., Thelwall, M., & Holyst, J. A. (2011b). Negative emotions boost user activity at BBC forum. *Physica A, 390*(16), 2936-2944.
6. 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.
7. Cornelius, R. R. (1996). *The science of emotion*. Upper Saddle River, NJ: Prentice Hall.
8. Dodds, P. S., & Danforth, C. M. (2010). Measuring the happiness of large-scale written expression: Songs, blogs, and presidents. *Journal of Happiness Studies, 11*(4), 441-456.
9. Fox, E. (2008). *Emotion science*. Basingstoke: Palgrave Macmillan.
10. Garas, A., Garcia, D., Skowron, M., & Schweitzer, F. (2012). Emotional persistence in online chatting communities. *Scientific Reports, 2,* article 402. doi: 10.1038/srep00402
11. Glorot, X., Bordes, A., & Bengio, Y. (2011). Domain adaptation for large-scale sentiment classification: A deep learning approach. *Proceedings of the 28th international conference on machine learning (ICML 2011).*

12. Gobron, S., Ahn, A., Silvestre, Q., Thalmann, D., Rank, S., Skowron, M., . . . Thelwall, M. (2011). An interdisciplinary VR-architecture for 3D chatting with non-verbal communication. *Proceedings of the Joint Virtual Reality Conference of EuroVR (EGVE 2011),* Nottingham, UK. 87-94.
13. Kramer, A. D. I. (2010). An unobtrusive behavioral model of "gross national happiness". *Proceedings of CHI 2010* (pp. 287-290). New York: ACM Press.
14. Krippendorff, K. (2004). *Content analysis: An introduction to its methodology*. Thousand Oaks, CA: Sage.
15. Kucuktunc, O., Cambazoglu, B. B., Weber, I., & Ferhatosmanoglu, H. (2012). A large-scale sentiment analysis for Yahoo! Answers. Paper presented at the *Web Search and Data Mining (WSDM2012),* Seattle, Washington. 633-642.
16. Liu, B. (2012). *Sentiment analysis and opinion mining*. New York: Morgan and Claypool.
17. 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)
18. Norman, G. J., Norris, C., Gollan, J., Ito, T., Hawkley, L., Larsen, J., . . . Berntson, G. G. (2011). Current emotion research in psychophysiology: The neurobiology of evaluative bivalence. *Emotion Review, 3*, 3349-359. doi: 10.1177/1754073911402403
19. Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval, 1*(1-2), 1-135.
20. Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up? Sentiment classification using machine learning techniques. *Proceedings of the conference on empirical methods in natural language processing* (pp. 79-86). Morristown, NJ: ACL.
21. Pennebaker, J., Mehl, M., & Niederhoffer, K. (2003). Psychological aspects of natural language use: Our words, our selves. *Annual Review of Psychology, 54*, 547-577.
22. Ponomareva, N., & Thelwall, M. (2012). Do neighbours help? an exploration of graph-based algorithms for cross-domain sentiment classification. *The 2012 Conference on Empirical Methods on Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL 2012),*
23. Skowron, M. (2010). Affect listeners. Acquisition of affective states by means of conversational systems. *Lecture Notes in Computer Science, 5967*, 169-181.
24. Skowron, M., Pirker, H., Rank, S., Paltoglou, G., Ahn, J., & Gobron, S. (2011). No peanuts! Affective cues for the virtual bartender. In R. C. Murray, & P. M. McCarthy (Eds.), *Proceedings of the Florida artificial intelligence research society conference (FLAIRS-24)* (pp. 117-122). Menlo Park, CA: AAAI Press.
25. Somasundaran, S., Namata, G., Wiebe, J., & Getoor, L. (2009). Supervised and unsupervised methods in employing discourse relations for improving opinion polarity classification. *Empirical methods in natural language processing (EMNLP 2009)* (pp. 170-179)
26. Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational Linguistics, 37*(2), 267-307.
27. Thelwall, M., & Buckley, K. (in press). Topic-based sentiment analysis for the social web: The role of mood and issue-related words. *Journal of the American Society for Information Science and Technology.*
28. 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.
29. 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.

30. 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.
31. 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.
32. Thelwall, M., Sud, P., & Vis, F. (2012). Commenting on YouTube videos: From Guatemalan rock to el big bang. *Journal of the American Society for Information Science and Technology, 63*(3), 616–629.
33. 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.
34. Wiebe, J., Wilson, T., Bruce, R., Bell, M., & Martin, M. (2004). Learning subjective language. *Computational Linguistics, 30*(3), 277-308.
35. Wilson, T., Wiebe, J., & Hwa, R. (2006). Recognizing strong and weak opinion clauses. *Computational Intelligence, 22*(2), 73-99.