Electronic Social Spaces, Web-Scrapers and Narrative Analysis: Problem Identification Tools for Web-based Counseling Support Services?

David A Banks, University of South Australia, Adelaide, Australia, david.banks@unisa.edu.au

Abstract

Web-based counseling services have grown since the inception of Internet, typically built around traditional counseling models where clients approach service providers. This paper considers an approach that is pro-active in finding individuals who may be at risk. The paper outlines the relationship between the content of text, in this instance located in electronic social spaces, and the underlying mental health of the writer. It is suggested that web scrapers could be utilized as a tool to trawl blogs to identify text patterns that may indicate that the writer may benefit from counseling support services. The paper only considers a possible technical approach and does not examine the substantial ethical issues that would arise from the implementation of such an approach in practice. The ethical aspects will be explored in a future paper.

1. Introduction

The Internet has offered a growing opportunity for providers of counseling services to reach parts of the community that are normally difficult to support. For example, members of drought-stricken communities are be placed under considerable stress but may have problems in accessing support services due to both lack of time and geographical location. Internet offers an attractive vehicle for providers of counseling services to reach such isolated individuals and offer support at times that suit the needs of clients. Equally there are members of communities that are city-based but are personally isolated due to social or financial circumstances. The 'net generation', used to communicating via Internet services, may also find that the existence of counseling services in that electronic space is a natural and acceptable way to seek support for any problems that they may have. Internet thus provides a wide range of individuals with a potential connection channel when they need support at some time in their lives.

The common feature of these new types of counseling services is that they are based upon a process that involves advertising of the services via Internet, response by potential clients and then development of a counseling relationship by some web-enabled mechanism such as email, chat or possibly a virtual environment such as Second Life. These are all effectively extensions of traditional counseling approaches that enable potential clients to establish contact with counseling services and then engage in a supportive process. The online nature of the process does raise some concerns including those of the processes management of 'remote' interaction, loss of some traditional cueing signals for the counselor, establishment of the credentials and trustworthiness of service provider, maintenance of confidentiality and verification of client identity. Many of these concerns can be, and indeed are being, addressed by reference to the traditional aims of counseling providers along with well established codes of conduct, although the latter may need to be re-evaluated or re-stated in online environments. Barak (2007) for example reports on the SAHAR system which is based upon the premise that there is a tendency for people to express many normally private feelings and concerns when communicating in cyber-environments and that Intenet could utilize this willingness to be open and disclose their 'inner' self. SAHAR operates only in the Hebrew language and offers informative articles and links to other relevant literature and web sites. It also lists details of support organizations, phone numbers and email addresses for those who need urgent help.

This paper considers a different, and potentially more contentious, possibility, where publicly available Internet narratives, for example blogs, could be regularly scanned to identify persons who may be at some kind of risk but have not sought any counseling support. In the approach explored in this paper public narratives would be scanned by means of automated web bots, with textual analysis being applied to those retrieved narratives to identify potential psychological risks, possibly followed by early intervention approaches to the writers of those narratives. There are clearly a number of significant ethical issues that arise from such an approach, but this paper concentrates mainly upon the technical practicality of such an approach rather than upon these ethical issues. The substantial ethical issues arising from such an approach will be explored in a future paper. In brief, the underpinning ethical premise for this paper is that if it can be demonstrated that practical and reliable web-based mechanisms can be developed to detect persons at risk of harm, and that follow-up mechanisms could be utilized to provide beneficial support then the idea is worthy of serious debate.

Internet can be a hostile environment for individuals who are experiencing, either knowingly or unknowingly, a crisis in their lives. If they are conscious of these stresses they may use the web as a vehicle to express their feelings via blogs and other social communication channels, but such expressions may or may not be received sensitively by others. In fact, there is evidence that harm can occur when such channels are utilized. Humphreys (2008), writing a personal investigative article for a UK newspaper notes that most of the conversation he came across were essentially no more than idle chatter, but in once instance he came across an individual expressing suicidal intentions. The response of several members of that particular space (Habbo) ridiculed him or told him to 'get on with it'. Bale (2007), reported the case of an individual who switched on his web cam and announced that he was going to commit suicide. He stood on a chair, made a hole in the ceiling, tied one end of a rope round his neck and the other end to an exposed beam. Some chat room users tried to tell him to stop, but others, either because they thought he was not serious or perhaps for other reasons, urged him to carry on. It should be noted that Ungoed-Thomas (2007) reported that this event took place in an 'insult' chatroom where people trade insults with each other and this may partially explain why the seriousness of the threat may not have been easily detected. When he did indeed step off the chair it was too late for anyone to organize help. Bale (2007) notes that this echoed a similar case in the US where an individual committed suicide on line by taking an overdose of drugs and alcohol. Again, although some of the 100 people in the chatroom at the time tried to prevent him from taking such action, others egged him on. Once an individual places themselves in this position it is difficult for those in a chat room to determine the seriousness of any threat and for appropriate authorities to be notified in time to prevent harm occurring. Had the individual who took his life contacted an online counseling service it is possible that actions could have been taken to prevent these tragedies. Barak (2007), for example, suggests that SAHAR has directly saved many lives since its inception, but this would appear to be the case only if the individual, or another concerned party, contacted the organization to initiate counseling support or other action. The argument in this paper is that if automated early detection systems could identify persons at risk this would allow appropriate support services to be mobilised before the individual reaches a state of potential self-harm or becomes a victim of individuals or groups who may see their situation as one to be exploited.

The remainder of this paper explores ways in which a variety of technologies can be utilized to offer the possibility of developing a proactive approach where persons at risk are identified by analysis of their written entries in electronic public spaces. Should entries suggest that the individual may be at risk, either because they have overtly stated that position or because analysis of the narrative entries suggests potential risk, an intervention strategy would be triggered. As noted earlier, this approach raises many ethical concerns, but the purpose of this particular paper is to demonstrate that such a system is technically viable. The next part of the paper considers the link between writing and psychopathology.

2. Text as an indicator of underlying states of mind

The study of text has been used for a considerable time as a vehicle for gaining deeper insight to a writers motivation or state of mind. Junghanel, Smyth and Santner (2008), for example, note that a considerable number of studies examining writing and psychopathology have successfully linked linguistic style to a variety of personality and health-related indices. Neuroticism has been found to correlate positively with the use of negative emotion words (eg, 'sad', 'anger', 'dislike') as well as first person pronouns (Pennebaker and King, 1999) and patients diagnosed with paranoia could be distinguished from other psychiatric disorders through the analysis linguistic factors (Oxman et al, 1982).

Narrative research can take a number of stances, from primarily qualitative through to primarily quantitative. It can be argued that language itself cannot be separated from the context in which it was expressed and therefore analysis has to be essentially qualitative, but an alternate view is that it is possible to count and statistically analyse words in expressed language in order to obtain understanding. Boals and Klein (2005) suggest there are three prevailing views relating to text analysis, supporting the qualitative and quantitative perspectives but also include a middle ground view based on the work of Bruner and Feldman (1996). For the purposes of this paper the end of the spectrum of views that is adopted is that which includes the quantitative counting and categorizing approaches. This is not to suggest that context is unimportant, because clearly in counseling situations context is vital to assist the counselor in understanding the situation and needs of the client. The quantitative end of the spectrum is being emphasized in this paper because it fits well with the technology being suggested here for identification of potential clients at risk. It would be anticipated that in practice any counseling contracts established as a result of this risk identification phase would make full use of the broad range of methodologies and tools as appropriate to each case.

Word count strategies assume that the words people use carry psychological information that transcend both their literal meaning and the surrounding context. Computer based approaches can be used to carry out the analysis using a variety of tools. One of these tools, Linguistic Inquiry and Word Count (LIWC) was developed by Pennebaker, Francis and Booth (Pennebaker et al, 2001) to support research into emotional writing. LIWC searches through text files, seeking over 2300 words or word stems which have been categorized into more than 70 hierarchically-organised linguistic dimensions. This approach allows examination of many dimensions, including pronoun use, positive affectivity, negative affectivity, cognitive processes and social processes. The analysis of the various categories can reveal psychopathological aspects of writing that relate to the emphasis given to self versus others, emotions such as anger or anxiety and to feelings of inclusion, exclusion, uncertainty and so on. LIWC has been used to successfully detect differences in writing on Internet homepages and message boards of pro-anorexics and recovering anorexics (Lyons, Mehl and Pennebaker, 2006), and differences in natural word use between psychiatric outpatients and nonclinical controls (Junghaenel, Smyth and Santner, 2008). Pennebaker, Mehl, and Niederhoffer (2003) report that a number of studies have demonstrated reliable links between computerized analysis of language and the classification of patients into diagnostic groups such as schizophrenia, depression, paranoia or other somatization disorders. They suggest that despite sometimes conflicting results, language can be an attractive as well as subtle diagnostic marker but that these links require more research to identify the underlying theory of these observations.

3. Analysis of publicly available texts

Publicly available narrative such as political speeches and poetry have been analysed by a number of authors and demonstrate that underlying aspects of mood, motivation or reaction to stress can be detected. Pennebaker, Mehl and Niederhoffer (2003) note that word choice for an individual is sufficiently consistent over time in a variety of situations to provide a measure of individual differences. For example, Pennebaker and Lay (2002) found that Mayor Rudy Guilani's use of first-person singular was greater during times of personal distress. They suggest that the patterns of pronoun usage may "serve as markers of emotional state, social identity, and cognitive styles" and may be predictive of psychological outcomes. Bucci and Freedman (1981) found that depressive states led individuals to use first person singular pronouns more often in speech, with a lack of second and third person pronouns. Stirman and Pennebaker (2001) found that archival study of the language of suicidal poets expressed more use of first person singular pronouns and less use of first person plural pronouns than nonsuicidal poets.

In a study of narratives used by pro-anorexics and recovering anorexics Lyons, et al (Lyons, Mehl and Pennebaker, 2004) carried out unobtrusive sampling of language from the Internet, specifically from on personal home pages and message boards. (Pro-anorexics are individuals who consider anorexia a legitimate lifestyle that they choose to have rather than an illness that they cannot control) Their LIWC based approach found that pro-anorexics manifested 'a more pronounced hedonic focus on positive emotions and the here and now, reduced level of cognitive processing and a lower degree of self-preoccupation' (p256) and that this could be identified from text in a reliable way.

Text analysis using such tools as LIWC has proved to be a useful way of identifying a number of characteristics of individuals and appears to be reliable in practice. The next part of the paper considers web scrapers as a tool for gathering narratives from public spaces for subsequent text analysis.

4. Narratives in the social spaces of the Web

'Blogging', the writing of personal journals, diaries or comments for posting to a common webpage, is a growing activity for many Internet users. One would have to question the veracity of some of the material, given that it could be either a true public diary or a piece of fiction. It is possible that some blogs and other materials written by individuals for public consumption may be acting as a vehicle for some individuals to project a different persona to their 'real' self, that is, to use the media as a public stage where they can perform in ways which they would not normally be able or allowed (Tantam, 2006). Equally, there may be a tendency to be more open and disclose more of the inner self due to dis-inhibition brought about by a sense of distance from other people in the electronic space. Any consideration of analysis of online materials in the form of blogs must therefore be framed within an appreciation of the possible motivations for the writing of those materials.

If we assume that a reasonable majority of the written material presented in blogs is genuine then we have an opportunity to analyse them using the narrative analysis techniques identified earlier in this paper. The link between written material and the state of mind of the writer should allow us to identify those pieces of writing that may have been generated by individuals who are some kind of stress or who have an actual or potential psychological problem.

Web scrapers have been with us for some time. These allow large number of Internet sites to be searched for target words and a piece of surrounding text to be retrieved along with details of the publication site for subsequent analysis. Details for building web scrapers are freely available on Internet and they do not appear to require extensive or high level programming skills. It should therefore be possible to build a system that searches for the presence of key words and phrases that could act as the trigger for more detailed analysis of specific blogs, possibly followed by some kind of intervention should that be felt to be appropriate.

Neilsen Buzzmatics has a number of products that have been developed to allow help clients to compare their product against others by listening in on 'unaided consumer conversations that take place on Internet forums, boards, Usenet newsgroups and blogs, providing timely understanding of the opinions and trends affecting your brands and the marketplace' (Nielsen BuzzMetrics, 2008). The suite of powerful tools that support BrandPulse include Relevance Detection, Classification, Phrase Mining, Sentiment Mining, Concept Mining, and Social Network Analysis. The Sentiment Mining tool is designed to 'identify polar expressions (positive and negative) in unstructured data … the Sentiment Miner can deliver a sense of 'emotionality' and opinion on topics vital to clients' needs'. For the purposes of this paper BlogPulse, also from Neilsen BuzzMetrics, was used to explore the type of material that could be expected to be retrieved from the Web through the use of targeted search systems. Blogpulse offers less functionality than BrandPulse, but has the attraction of being free, and this was used to carry out some tentative tests of the idea explored in this paper. BlogPulse allows a user enter words or phrases and will then search blogs to identify those target phrases and to generate a graph of the frequency of use of those words over a prescribed period. The mouse pointer can be moved along the resulting graph to obtain a link to the source. The key word used in this paper was 'suicide'. As would be expected the graph showed periodic peaks that coincided with major suicide events around the world but also revealed some 'hits' that suggest that the approach discussed in this paper is viable. Typical broad results, in the form of the fragments returned as a result of the search, are shown below:

"I have an idea – lets rule the universe together! The mass suicide of the aliens at the end seemed a bit random …"
"It is imperative that an end is brought to suicide bombings …"
"First we were told that Bhutto was shot twice, once in the head and once in the shoulder, before the suicide bomber blew himself up …"

Hits that are more relevant from the first ten on the list are:

"A cousin committed suicide on Christmas Eve …

"I though seriously about a tattoo, but didn't. My ex-stepdad committed suicide. My mom was a wreck. We experienced a second suicide within my family"

Each listed source can be examined in full by going to the site that the fragment was drawn from and the material would thus lend itself to detailed analysis using computer-based tools such as LIWC. An example of one full text, dated 20-03-2008, retrieved from this trial search is shown below, from a LiveJournal™ entry:

"Alright here it is.
SAT scores came in. Grades are going out. Life goes on. How much of these things can I not care about? Sure, people say grades don't dictate how well you live your life, you'll still live with shitty SAT scores, just keep on going and you'll get to wherever you want some day.

i don't know. When you're contemplating what personalized suicide note you should give to all your friends and enemies, you probably know your life sucks. My life doesn't suck, really. I'm pretty average. I just don't want to be.

See, I fit in perfectly with everyone who I don't want to be with. My life is trodden over and was spent kneeling and kissing and sucking up to people who I considered to be above average and successful. But really...how do I even compare up to them? I can't keep a steady conversation going with whatever is going on in their lives/the world because I don't know enough of either. I can't keep my grades up to nearly match theirs. I'm not strong or as built/athletic as they are. I don't even have a job anymore. Fuck. I even suck at video games.

There's a point where I can smile and shrug off 'life changing' issues (something I've been doing successfully for a long time) but here, when I got the shit beat out of me by someone I didn't know, then the beat shit out of me by another person in the beginning of the fucking tournement bracket...I've been spending my whole life striving for nothing.

I'm not good at anything, I can't think through things like other people. I know I'm slow. I can't think of clever things to say all the time. A lot of

the time I just can't grasp words to explain how I feel.

I think I just exist to break peoples hearts and to waste bandwidth."

No attempt has been made to carry out an analysis of this extract as the intention is purely to demonstrate that the use of web scrapers followed by analysis of retrieved materials to determine risk factors is a viable process and may have some value for mental health support organisations. If the sentiment expressed in this example is genuine, then clearly this is individual would benefit from a supportive intervention to help them build their self-esteem. If a skilled counselor could establish a dialogue with this individual they may be able to provide immediate support and identify deeper issues that may need to be addressed. The technology is thus only a small part of the overall process discussed in this paper, providing a tool to identify potential personas at risk. Traditional counseling support system would be needed beyond that initial identification process.

Any such material captured through blogs is the result of complex human processes and the underlying motivations for such expressions of feeling will be clearly themselves be complex. Although there appears to be substantial and credible evidence that individual written narratives do reflect the mental states of individuals there are other influences that may produce similar signals suggestive of stress but that may only of a transient nature. Joiner, Hollar and Van Orden (2006), for example, found that some suicide risks can be reduced if individuals feel to be part of a collective community, particularly when that community is demonstrating success. They found that suicide rates could be strongly associated with the performance of sports teams – when teams were performing well their supporters experienced an overall 'pulling together' effect and suicide rates were lower than when the teams performed poorly. Other influences upon the interpretation of web narratives is that the act of writing is itself used as a therapeutic tool and a blog may provide reflective tool that offers an opportunity for individuals to work through their issues. Baikie et al (2007) report that positive physical and/or psychological benefits can arise from expressive writing, although no theoretical understanding of why this should occur seem to be available. Kacewicz, Slatcher, and Pennebaker (2008) suggest that the act of writing, particularly about ones-self, requires reflection and acknowledgement of the underlying emotions and that this may process may be beneficial for some individuals with particular mental health issues. Too early an intervention may thus interfere with, and possibly hinder, a self-managed process of understanding and dealing with problems. On the other hand, as identified earlier in the paper, the 'blogosphere' is inhabited by a wide range of individuals not all of whom may offer support and some may in fact, for their own reasons, use the opportunity to cause damage to others.

5. Conclusion

The purpose of this paper is to articulate the link between publicly available written material in the form of blogs and ways in which this material could be automatically analyzed as a means of identifying individuals who may potentially benefit from an early intervention by counseling or other appropriate mental health practitioners.

The web provides an environment where troubled souls may find some solace by projecting another image that is stronger than their own or may seek support from an anonymous audience when in personal crisis. Those that project false images of themselves may deepen their isolation by amplifying the acceptance of their 'other self' by others at the cost of heightening the contrast with their 'real self'. In both situations those that are vulnerable are potentially exposed to risk of personal harm in some circumstances. At the same time that there is growth in the number of troubled individuals we are also seeing pressure on the counseling resources that could support them.

This paper has explored the idea that web-based tools and systems may provide opportunities for an automated early identification of persons at possible risk. Intervention approaches to pro-actively offer support for such individuals who may be at risk could be developed by appropriate support services. The support mechanisms offered may be web-based using email, chat or other communication channels, or they could take place in traditional face-to-face settings. Such pro-active interventions may only form a small extra part of much broader web-based counseling developments and there are substantial technical, legal and ethical problems to be taken into account before such an approach could be used in practice. It is intended to examine the ethical issues in a further paper.

6. References

Baikie, K. A., Wilhelm, K., Johnson, B., Boskovic, M., Wedgwood, L., Finch, A. and Huon, G. 2007, "Expressive writing for high-risk drug dependent patients in a primary care clinic: A pilot study", *Harm Reduction Journal*, BioMed Central

Bale, J. 2007, **"**Get on with it, said net audience as man hanged himself on webcam", *The Times*, March 24th, UK

Barak, A. 2005, "Emotional support and suicide prevention through the Internet: A field project report", *Computers in Human Behavior*, 23, pp

971-984
BlogPulse, http://www.blogpulse.com, accessed 22/02/2008
Boals, A., and Klein, K. 2005, "Word use in Emotional Narratives about Failed Romantic Relationships and Subsequent Mental Health", *Journal of Language and Social Psychology*, 24, pp252-268
Joiner, T. E., Hollar, D., Van Orden, K. 2006, "On Buckeyes, Gators, SuperBowl Sunday, and the Miracle on Ice: "Pulling Together" is Associated with Lower Suicide Rates", *Journal of Social and Clinical Psychology*, Vol 25, No. 2, pp 179-195
Junghaenel, D. U., Smyth, J. M. and Santner, L. 2008, "Linguistic Dimensions of Psychopathology: A Quantitative Analysis", *Journal of Social and Clinical Psychology*, Vol 27, No 1, pp 36-55
Kacewicz, E., Slatcher, R. B. and Pennebaker, J.W. 2008, "Expressive Writing: An Alternative to Traditional Methods", *Handbook of Low-Cost Interventions to Promote Physical and Mental Health: Theory, Research and Practice*
Lyons, E. J., Mehl, M. R. and Pennebaker, J. W. 2006, *Journal of Psychosomatic Research*, 60, pp 253-256
Neilsen Buzzmatics, 2008, http://neisenbuzzmatics.com, accessed 23/03/2008
Pennebaker J, W., Francis M. E., Booth R. J. 2001, *Linguistic Inquiry and Word Count* (*LIWC*)*:LIWC 2001*. Mahwah, NJ: Erlbaum
Pennebaker, J.W. and Lay,T.C. 2002, "Language use and personality during crises: Analyses of Mayor Rudolph Giulani's press conferences". *Journal of Research in Personality*, *36*, pp 271-282.
Pennebaker, J. W., Mehl, M. R. and Niederhoffer, K. G. 2003, "Psychological Aspects of Natural Language Use: Our Words, Our Selves", *Annu. Rev. Psychol*., 54, pp 547-577
Tantam, D. 2006, "Opportunities and risks in e-therapy", *Advances in Psychiatric Treatment*, Vol 12, 2006, pp 368-374
Ungoed-Thomas, J. "Police hunt chatroom users over web suicide 'goading'", *Sunday Times*, March 25, UK