Crowdsourced Real-world Sensing: Sentiment Analysis and the Real-Time Web

Adam Bermingham and Alan F. Smeaton

CLARITY: Centre for Sensor Web Technologies,
School of Computing,
Dublin City University
{abermingham, asmeaton}@computing.dcu.ie

Abstract. The advent of the real-time web is proving both challenging and at the same time disruptive for a number of areas of research, notably information retrieval and web data mining. As an area of research reaching maturity, sentiment analysis offers a promising direction for modelling the text content available in real-time streams. This paper reviews the real-time web as a new area of focus for sentiment analysis and discusses the motivations and challenges behind such a direction.

Keywords: sentiment analysis, real-time web, microblog

1 Introduction

In the last 10 years, user-generated content has come to dominate a large portion of the web. Reviews, blogs, social networks, discussion forums and wikis are all familiar concepts to the average Internet user. User-generated content has now earned respect as a credible source for exploring both factual and subjective information. This has inspired research in the area of automatic *sentiment analysis*: methods for automatic detection of negative and positive emotions, opinions and other evaluations in text.

The *real-time web* refers to the portion of the web where information is available shortly after it is created and where it is connected in some way with events that are happening in the real world either at, or close to that time. In terms of user-generated content, the information takes the form of blog posts, microblog posts, news feeds and social network content amongst others. This content is often reactionary in nature, disseminating news of real-world events in real-time and expressing associated opinion and commentary. Just as events in the real-world can happen at specific times and are scheduled, or can be unpredicted and occur spontaneously, so too does user-generated content have a prominent time component. Examples of scheduled real-world events would be sporting contests and TV programs and spontanous real-word events would include riots and civil disturbances.

The microblogging service, Twitter, is a good example of information making up the real-time web. Twitter allows users to publish short text messages and these messages then appear in their followers' feeds and may appear in searches. Twitter users write about a wide variety of topics including both scheduled and spontaneous real-world and real-time events. The diversity of content, the large volume and the availability of data mean that Twitter provides us with a unique opportunity to mine sentiment in real-time in a way not possible before. Throughout this paper we use Twitter as a case study for sentiment in the real-time web.

The task of applying sentiment analysis to the real-time web however has a number of challenges. Due to the dynamic nature of the real-time web, topics of interest are constantly evolving. There are also infrastructural challenges as static indexes and latency in computation are no longer acceptable. This also poses difficulties for experimental methodology as methods for evaluating techniques for real-time knowledge discovery are not well established.

This paper offers a review of sentiment analysis in the real-time web as a future direction for research. The rest of this paper is laid out as follows: in Section 2, background to sentiment analysis is presented followed by a motivation for pursuing this line of research in Section 3. Challenges are discussed in Sections 4 and we conclude in Section 5.

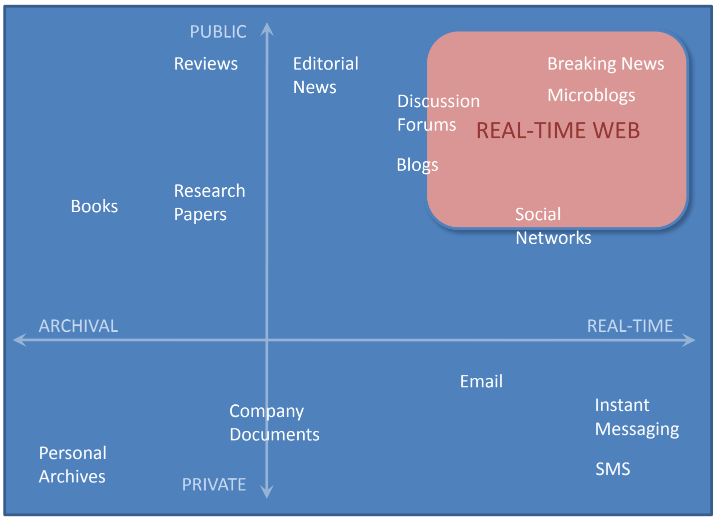

Fig. 1. A conceptualization of the the real-time web in terms of digital content.

2 Background

Much sentiment analysis research concerns the study of static data. A number of publicly available collections have been widely used in research such as Pang and Lee's movie review corpus [6], Wilson *at al.*'s MPQA corpus [10], Blogs06 from the TREC Blog Track [4] and the data provided as part of the NTCIR opinion finding task [8] to name a few of the more common examples. These corpora may all be thought of as static corpora: all of the topics are predefined and the labels are independent of any temporal aspect. This has many advantages, the replicability of experiments being one. However, in the real-time web, documents, topics and sentiment all have an inherent temporal component.

Recently there have been research works based on Twitter. Bollen *et al.* [1] modelled trends in mood on Twitter in the 6 dimensions defined by the psychometric test, the *Profile of Mood States* (POMS): *tension*, *depression*, *fatigue*, *vigour*, *anger* and *confusion*. They developed a term-based emotional rating system by extending the 65 adjectives defined in POMS to a lexicon of 793 terms. They then use the variance of these terms in Twitter posts over a series of 153 days to model the trends in mood along the 6 dimensions. They found that sentiment on Twitter correlated with real-world values such as stock prices and coincides with events. They conclude:

> "events in the social, political, cultural and economical sphere do have a significant, immediate and highly specific effect on the various dimensions of public mood."

Similarly, Diakapolous and Shamma analysed sentiment in Twitter posts to characterise the reaction to various issues in a US Presidential election debate in 2008 [2]. Rather than use an automated sentiment analysis approach, they crowdsourced the annotations using Amazon Mechanical Turk (AMT). The AMT annotators were asked to annotate documents which discuss the debate as *positive*, *negative*, *mixed* or *other* towards each of the presidentatial candidates. Diakopolous notes significant shift in sentiment as the debate moves between speakers and topics. The also suggest that by correlating the positive and negative sentiment they can identify controversial issues, though they note that this requires further investigation. They conclude by offering two caveats surrounding modelling sentiment in real-time streams: (i) the relationship between sentiment and a real-world occurence is inferred based on document timestamp and relevant terms and is not necessarily accurate in all cases and (ii) the authors of the documents in the stream are not necessarily representative of the wider population.

These works support the assumption that the sentiment in the document stream is indicative of people's reaction to real-world events. This is encouraging as it demonstrates that real-time social content, such as Twitter, is a valid data source for gauging public sentiment.

Another promising avenue of research is in market research. One such work is Jansen *et al.* who studied the Word-Of-Mouth effect on Twitter, focusing on how and why positive and negative sentiment towards brands spreads on Twitter [3]. They use supervised learning to classify twitter posts which mention brands for sentiment towards that brand using n-gram features. Interestingly, they find automated sentiment analysis accuracy comparable to manual classification. Of the documents they analyse, 19% contain a reference to a product, company or service and of these, 20% contained sentiment towards that product, company or service. They also observe large temporal swings in sentiment and suggest that marketing companies are required to continually monitor their streams of documents over time. Again the results demonstrate that Twitter data provides a means to sense the collective sentiment towards topics of interest.

These exploratory works hint at the potential of researching automatic techniques to model the sentiment in the real-time web. There remains much work to be done to explore fully the possibilities presented by the real-time web.

3 Motivation and Applications

In applying sentiment analysis to the real-time web, and in specific user-generated content, we are in essence crowd-sourcing our sensing of the real world in real-time. The online conversation becomes a sea of data from which we can infer sentiment and extrapolate information about the real-world around us. This is not something that has been possible until now in any meaningful way and so we are presented with a unique avenue for research.

For some time there have been methods of near-instantaneous computer-mediated communication. Instant messaging (IM) and text messaging on mobile phones (SMS) are two such examples. Each of these types of communication however are intrinsically private and obtaining and publishing datasets based on the private correspondence of users is problematic at best. The public nature of the Internet means that no such privacy restriction exists in terms of mining the information in online content, real-time or otherwise. The standardised way in which this content is made available not only encourages developers and users to better use the content, but also us as researchers to efficiently construct datasets and data streams to be used for study.

The recent growth in the volume in the real-time web, specifically on Twitter, is staggering. At least one website has recently measured the rate of Twitter posts being published is 2 billion per month, or 64 million per day, and increasing[1]. This is undoubtedly a large volume of information to analyse, even given the short length of twitter posts. But what portion of this deluge are relevant to a given topic interest? In the recent Soccer World Cup in South Africa, even the early matches saw activity in the region of hundreds of thousands of tweets per match. Similar activity was seen during the NBA play-offs. High levels of activity are also seen in relation to unfolding news stories and live television.

Thus the need for automated analytical and aggregation techniques is clear. Search on Twitter[2] is dominated by inverse chronologically ordered results, fil-

[1] http://royal.pingdom.com/2010/06/08/twitter-now-2-billion-tweets-per-month/

[2] http://search.twitter.com

tered by keyword. In this model, the assumption is that recency is the single most important measure of relevance. With many relevant documents being produced, there will be many more before a user has time to finish reading the search results. This simple model does not scale well. The problem of search in the real-time web is still an unsolved problem. Perhaps real-time streams of user generated content are destined to be passively observed rather than actively searched. The problem definition and methodologies are still in flux. By enriching the documents with sentiment information, the opportunity is there to employ more sophisticated methods to help users find useful information. For example, ensuring a level of diversity and representativeness of sentiment in the results list.

As well as helping users find documents of interest, there is also value in being able to determine the aggregate sentiment in a real-time stream of documents. Being able to quantify sentiment for a given topic over time permits us to use sentiment as we would a stream of any other source of data: stock prices, sensor data. A real-time sentiment trend then allows us to find events that trigger deviations in sentiment and to integrate this with other data feeds both from the online and offline world. This type of aggregation and high performance analytics is of obvious benefit to many areas of industry, government and research.

Sentiment analysis is an area of research reaching maturity (see [7] for a detailed history of the field). There are now established methodologies, in particular for machine learning techniques, for obtaining accuracies comparable with the traditionally easier task of topical classification. It is the intersection between (i) the abundance and availability of data, (ii) the maturity and of sentiment analysis as an area of research and (iii) the dearth of research into sentiment-based strategies for real-time information analysis that motivate this area of research.

4 Challenges

The primary challenge in the real-time web is understanding the information needs and interaction patterns. We have already seen how real-time services such as Twitter are both a disruptive and a challenging and opportunistic technology. As of now it is unclear what are the common interaction patterns and perhaps more importantly, which ones will prove to be beneficial as the technology matures to a significant degree of productivity. Again taking Twitter as an example, perhaps search will prove to be the most valuable way of interacting with Twitter, as it has been for the traditional web. On the other hand, perhaps the real value is being able to navigate the people you personally follow, a more social network oriented perspective. Perhaps focus will shift from journal style content to more topical content, as has arguably happened in the blogosphere. Perhaps the real-time web in the context of an event behaves quite differently to that at another arbitrary time. Or most likely, perhaps the overall picture is one which is more complex and which warrants careful thought and consideration.

From a sentiment analysis point of view, the real-time web means pushing sentiment analysis beyond review classification. Review classification serves as a good constrained experiment to evaluate sentiment analysis techniques but is less relevant in a real-world ad-hoc domain. In review classification often the topics are homogenous and there is little or no topic drift in the documents. Twitter by it's nature is dynamic and unpredictable, even the sentiment topics of interest may themselves may only become apparent as a real-world event unfolds and not be conceived beforehand. These types of ad-hoc scenarios can be troublesome. Add to this the variability in topic nature and the temporal dependency of training data used and there are a lot of challenges in approaching the accuracy enjoyed in classifying reviews.

Computational efficiency is also a consideration. Some of the higher performing sentiment analysis systems (for example [5]) have relied on computationally expensive feature extraction techniques such as parsing and dependency extraction to achieve their results. Without extensive computing resources, this would likely be unfeasible for the forseeable future with a required throughput of many documents per second. These problems can be mitigated by sampling strategies or, for aggregate sentiment, by allowing for a latency or less granular sentiment reading.

As an informal communication platform, Twitter exhibits characteristics of noisy text. Twitter posts often contain spelling errors, grammatical contractions, non-standard punctuation, emoticons etc. Tagliamonte analysed English language use in Instant Messenger (IM) by teenagers and adolescents and found that although the text exhibited features of noisy text, these patterns were not prevalent in his older participants and that this type of text was not as common as construed in the media [9]. In any case, with sufficient training data n-gram models (or extensions thereof) should robustly handle arbitrary tokens. This could degrade performance of approaches who rely on parsing to extract features from the text. Such approaches may benefit from a step of language normalisation where the text is amended, either heuristically or using a machine translation approach, to a more standard form. However, non-standard language usage may even prove to be beneficial to learning approaches. Often an author may do this to express themselves in a more concise way where they may be constrained by length of by the modality they are using for input. For example, the following punctuation sequences all add tone to the content of a document: "...", ":-)", "?!", "!!!!".

Evaluating methods for sentiment analysis in real-time also poses a significant challenge for research methodologies. It is likely that a true measure of the effectiveness of improving information discovery using sentiment analysis is not possible to determine outside of real-time. In real-time, perhaps our only option is the expensive task of evaluating users' interactions with systems by inferring effectiveness from their use of the system. Another option is to prompt users for real-time system feedback. This type of evaluation can borrow from the field of user interface evaluation where contrasting methods of interaction are evaluated. In either case, this is not as scalable or as reproducible as some common evaluation methodologies such as multiple fold cross validation in machine learning or the Cranfield model for evaluating information retrieval systems.

These challenges collectively are not insurmountable and form a number of interesting research questions for the field of sentiment analysis to pursue.

5 Conclusion

The real-time web has disrupted and challenged our methods for managing information retrieval and knowledge discovery on the web. Our research methodologies for analysing static corpora, independent of time, need to be rethought and we need to adapt our infrastructures for dealing with such time dependent information. At its core this is an exciting time for user-generated content. In a short space of time, the Internet has a become a place where information and thoughts about a vast array of topics can be gathered nearly instantaneously. Without a well-designed method for organising this information and allowing users to access and leverage it, the data becomes redundant, and we are not able to take advantage of the opportunities presented by the advent of the real-time web.

Acknowledgments

This work is supported by Science Foundation Ireland under grant 07/CE/I1147

References

1. J. Bollen, A. Pepe, and H. Mao. Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. *CoRR*, abs/0911.1583, 2009.
2. N. A. Diakopoulos and D. A. Shamma. Characterizing debate performance via aggregated Twitter sentiment. In *Conference on Human Factors in Computing Systems (CHI 2010)*, 2010.
3. B. Jansen, M. Zhang, K. Sobel, and A. Chowdury. Twitter power: Tweets as electronic word of mouth. *Journal of the American Society for Information Science and Technology*, 2009.
4. C. Macdonald and I. Ounis. The TREC Blogs06 collection : Creating and analysing a blog test collection. Technical report, University of Glasgow, Department of Computing Science, 2006.
5. S. Matsumoto, H. Takamura, and M. Okumura. Sentiment classification using word sub-sequences and dependency sub-trees. In *Proceedings of PAKDD'05, the 9th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining*, 2005.
6. B. Pang and L. Lee. A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts. In *ACL '04: Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, page 271, Morristown, NJ, USA, 2004. Association for Computational Linguistics.
7. B. Pang and L. Lee. Opinion mining and sentiment analysis. *Foundation and Trends in Information Retrieval*, 2(1-2):1–135, 2008.
8. Y. Seki, D. K. Evans, L. Ku, L. Sun, H. Chen, and N. Kando. Overview of multilingual opinion analysis task at NTCIR-7. 2008.

9. S. A. Tagliamonte and D. Denis. LINGUISTIC RUIN? LOL! INSTANT MESSAGING AND TEEN LANGUAGE. *American Speech*, 83(1):3–34, 2008.
10. T. Wilson, J. Wiebe, and P. Hoffmann. Recognizing contextual polarity in phrase-level sentiment analysis. *Proceedings of the 2005 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 347–354, 2005.