An Evaluation of the Role of Sentiment in Second Screen Microblog Search Tasks

Adam Bermingham and **Alan F. Smeaton**
CLARITY: Centre for Sensor Web Technologies
School of Computing, Dublin City University

Abstract

The recent prominence of the real-time web is proving both challenging and disruptive for information retrieval and web data mining research. User-generated content on the real-time web is perhaps best epitomised by content on microblogging platforms, such as Twitter. Given the substantial quantity of microblog posts that may be relevant to a user's query at a point in time, automated methods are required to sift through this information. Sentiment analysis offers a promising direction for modelling microblog content. We build and evaluate a sentiment-based filtering system using real-time user studies. We find a significant role played by sentiment in the search scenarios, observing detrimental effects in filtering out certain sentiment types. We make a series of observations regarding associations between document-level sentiment and user feedback, including associations with user profile attributes, and users' prior topic sentiment.

Introduction

Recently, sentiment analysis classification accuracies have become comparable with the traditionally easier task of topical classification. Given sentiment analysis's capabilities and limitations, we endeavour to demonstrate its benefit in application areas. For the task of search through user-generated content, analysis of query logs have shown that information needs frequently have a subjective component, for example in blog search (Mishne 2007). Our observations are that real-time events tend to be polarising and content is partisan (e.g. politics, sports) or critical (e.g. television). We argue that in the real-time social web, users' real-time needs have prominent sentiment components.

In this work, we describe a system we have developed for allowing users to view a stream of real-time content from the microblogging service, Twitter[1], while observing events. Our work has two aims: (i) to present a system and methodology for evaluating the role of sentiment in real-time microblog search, and (ii) to explore the relationship between sentiment and users in a real-time context.

Background and Related Work

Microblog search is perhaps most closely related to blog search. Two primary categories of blog search query are *concept* and *context* (Mishne and de Rijke 2006). Whereas *concept* queries concern a topic or area of interest, *context* queries aim to find commentary on real-world entities such as products or public figures. Mishne and de Rijke remark how this significantly differs from web search information needs, which are described as *informational*, *navigational*, or *transactional* (Broder 2002).

The prevalance of *context* queries in blog search has inspired much work on opinion-based search, for example in the Blog Track at TREC (Macdonald et al. 2010), and the Multilingual Opinion Analysis Task at NTCIR (Seki et al. 2010). One conclusion from the TREC Blog Track, was that due to the inherently opinionated nature of blog content, strong ad-hoc retrieval systems with no sentiment-specific techniques performed well on the opinion-finding task (Ounis, Macdonald, and Soboroff 2008), suggesting sentiment plays only a minor role in blog search. We argue that sentiment is more important in microblog search.

There are two types of microblog search query we might consider: (i) *Ad-hoc:* A user has an instantaneous information need at a point in time, and desires a single set of documents, and (ii) *Persistent:* A user wishes to state an information need, and receive documents which satisfy this need, as and when they become available. The former type of query is perhaps easier to formulate in a traditional information retrieval evaluation. Such a task has recently been run at the TREC Microblog Track[2]. The focus of our evaluation, *persistent* queries, allow people to track live events such as television programmes, breaking news stories and sports, as the event is unfolding. A common form of this is following an event on the social web, while also watching the event on television, known as *second-screen viewing* (Deller 2011).

A *persistent* microblog query may be thought of as the user expressing a wish to be shown documents which provide them with additional contextual information and commentary related to the query over time. Just like blog search, this does not conform to the notions of information need which epitomise web search, and is more like an informa-

[1] http://www.twitter.com

[2] https://sites.google.com/site/trecmicroblogtrack/

tion filtering system (Belkin and Croft 1992).

One work that has tackled microblog search improves retrieval performance over a Boolean search recency-ranked baseline using query expansion and quality indicators to extend a language model information retrieval approach to microblogs (Massoudi et al. 2011). Another work also uses language models to tackle microblog search (Efron and Golovchinksy 2011). In other research, Efron identifies the primary information retrieval tasks in microblogs as question answering, and what we refer to as *ad hoc* queries (Efron 2011). In this work, emphasis is placed on the prevalence of named entities as topics, and the implications of the presence of temporal context and meta-information. Both this and the previous work from Massoudi *et al.* treat the information need as instantaneous, and derive their methodology from traditional, static information retrieval evaluation.

A recent comparison of web search and microblog search through query log analysis notes that Twitter queries tend to be shorter than web queries, and are likely to be related to hashtags (Teevan, Ramage, and Morris 2011). The authors find, "*Twitter search is used to monitor content, while Web search is used to develop and learn about a topic.*" Some recent works tackle microblog search as a filtering problem, filtering tweets into general categories such as *news* and *events* (Sriram et al. 2010) or using social information to generate user profiles (Churchill, Liodakis, and Ye 2010). This area is, however, largely unexplored, perhaps due to the poorly understood information needs of persistent queries, and difficulties in evaluating such.

Experimental Design

Evaluating real-time microblog search with a static corpus evaluation is problematic. Relevance measures may not adequately discriminate in a large set of relevant documents. Also, static evaluations rely on the objective judgments of assessors, which for real-time search is subject to hindsight bias. A third problem is that objective judgments do not account for a user's internal knowledge or outlook. One recent review of search in microblogs concluded, "*...naturalistic and behavioral methods of assessing system performance will no doubt have a large impact on future research, as we work to make our studies both realistic and generalizable.*" (Efron 2011) This motivated our decision to conduct our evaluation with controlled user studies.

We characterised three aspects of persistent microblog search with respect to a given topic: (i) document sentiment, (ii) stream sentiment distribution and (iii) user sentiment, constituting our independent variables. We capture information for each user concerning their task familiarity and demographics and evaluate these as secondary, independent variables. Using a repeated measures experimental design, we exposed participants to various sentiment conditions, and recorded their feedback (our dependent variable). We adapted a familiar web interaction metaphor, *thumbs up* and *thumbs down*, and instructed participants to approach liking and disliking a document as they would if they enountered it in their normal Internet use. We also recorded periodic stream-level feedback by prompting users for a rating from 1 (*poor*) to 7 (*excellent*).

Participant Profile		XF	GE11
Age	≥25	17	18
	<25	18	3
Task Familiarity	slightly or not familiar	19	9
	somewhat or more familiar	16	12
Gender	female	12	10
	male	23	11
Prior Sentiment	positive	16	4
	negative	9	4
	neutral	7	9
	unfamiliar	3	4

Table 1: Participant sample sizes for profile attributes

	XF	%	GE11	%
positive	2,131	30.8	884	12.2
negative	3,640	52.7	2,716	37.6
neutral	843	12.2	3,628	50.2
mixed	296	4.3	153	2.1
Total	6,910		7,381	

Table 2: Labelled training documents for sentiment

We chose two real-time topics: the final of a singing competition on ITV, the X Factor (11th and 12th of December, 2010), and the Leaders' Debate during the Irish General Election (14th February, 2011). We recruited participants through the university staff and students (see Table 1). Participants observed the topic event live on a shared screen. When the event began, the system allocated each participant a random sentiment algorithm. At intervals of 15 minutes, the system prompted participants to provide stream-level feedback and the algorithms were rotated.

For a set of documents, T_{Up} is thumbs up frequency, and T_{Dn} is thumbs down frequency. T_{Net} is $T_{Up} - T_{Dn}$. Our experiment has four experimental conditions: (i) positive documents only (`pos`), (ii) negative documents only (`neg`), (iii) positive and negative documents only (`posneg`) and (iv) random sampling (`control`). We use the general linear model for repeated measures to compare the feedback distribution under the four conditions. In addition, the control algorithm is compared to the others using a paired, two-tailed, t-test. Using the general linear model also enables us to look at between-subjects main interaction, i.e. if there is a difference in the main effect, that corresponds to attribute differences between participants. Where we examine categorical associations, statistical significance is noted according to Pearson's chi-squared test.

The sentiment targets we use for the X Factor are the judges and contestants. Similarly, for the election our sentiment targets are the parties and their leaders. Our annotators labelled documents with respect to these sentiment targets (see Table 2). For the X Factor, we observed an agreement of 0.78 (Krippendorff's α) for 3 classes: positive, negative and neutral. For the election, the agreement was lower at 0.48, possibly reflecting a more difficult annotation task.

At search-time, we consider a positive document to be one which refers positively to each sentiment target that it mentions, and a negative document to be one which refers

	Feedback	posneg	pos	neg	control
XF	Overall*	4.21	4.06	4.61	4.35
	T_{Up} Rate**	0.19*	0.15**	0.24	0.23
	T_{Dn} Rate*	0.21	0.22*	0.17	0.17
	T_{Net}**	-0.01**	-0.07**	0.07	0.05
GE11	Overall	4.05	4.14	4.24	4.38
	T_{Up} Rate*	0.31	0.26*	0.32	0.32
	T_{Dn} Rate	0.13	0.13	0.12	0.13
	T_{Net}	0.17	0.13	0.20	0.18

Table 3: Mean feedback for sentiment filtering algorithms

negatively to each of the sentiment targets it mentions. A neutral document is then one which mentions one or more sentiment targets but does not contain sentiment towards those targets. During the X Factor we used two binary SVM classifiers, one for positive, one for negative. Using 10-fold cross validation, the accuracies for these classifiers on the training data were 82.47% and 75.57%, respectively. For the Leaders' Debate we used a three-way Adaboost multinomial naive Bayes classifer (positive, negative, neutral). This classifier as well as our feature vector and annotation methodology are described in our earlier work (Bermingham and Smeaton 2011).

Results and Discussion

In this section we present and discuss the results of our experiments.

Algorithm Sentiment

The average overall ratings for the sentiment filtering algorithms were slightly better than the midpoint of the 7-point scale (see Table 3). The streams, which upweight positive documents (`posneg`, `pos`), received lower ratings than those that do not (`neg`, `control`). However, this difference is only significant for the X Factor ($p < 0.001$).

The feedback for the Leaders' Debate was far more positive with a T_{Up} more than twice the T_{Dn}, whereas for the X Factor, T_{Up} was similar to T_{Dn}. For T_{Up}, we see a significant difference between the algorithms, with the `pos` algorithm again performing lowest for both the X Factor and the Leaders' Debate. Comparing algorithms to the `control` algorithm, it is the `pos` algorithm once more that demonstrates a significantly worse response for the X Factor T_{Up} and T_{Net} ($p < 0.001$), and T_{Dn} ($p < 0.05$). This pattern is also present for T_{Up} for the Leaders' Debate ($p < 0.05$). In both experiments, a document in the `pos` stream was considerably less likely to receive *thumbs up* feedback than in the `neg` or `control` stream. Also, the `posneg` algorithm performs significantly worse than the `control` for T_{Up} and T_{Net}, although the effect size is smaller.

The only significant differences in feedback according to the between-subjects main effect was for participants in different age groups for the X Factor study ($p < 0.05$). The algorithm ratings were higher for participants under the age of 25 for both document-level and stream-level feedback.

The patterns in sentiment are much more salient in terms of the X Factor than the Leaders' Debate. We did however have fewer participants for the debate and the event itself was shorter, so it is possible that some of our inconclusive results are subject to type II error.

		Thumbs Up	Thumbs Down
XF	positive	-0.31**	0.18**
	negative	0.3**	-0.16**
	neutral	0	-0.11*
GE11	positive	-0.08	0.11*
	negative	0.02	-0.04
	neutral	0.07	-0.08

Table 4: *Thumbs up* and *thumbs down* feedback log odds per document sentiment type

In modifying the sentiment in the stream we are possibly introducing negative effects other than upweighting an undesirable document sentiment type. For one, we are obscuring the true distribution of sentiment from the user. We are also potentially limiting their exposure to documents of other sentiment types. On reflection, these factors may have contributed to the strong performance of the `control` algorithm.

Document Sentiment

In total there were 55,314 documents presented to users during the X Factor and 7,509 documents presented to users during the Leaders' Debate.

For the X Factor, negative documents were twice as likely to receive *thumbs up* ($p < 0.001$), whereas positive documents on the other hand were just half as likely to receive *thumbs up* ($p < 0.001$) (see Table 4). For the Leaders' Debate we observed no statistically significant sentiment-feedback dependencies. We saw significant *thumbs down* patterns for positive and negative documents for the X Factor ($p < 0.001$); positive documents were 52% more likely to receive a *thumbs down* annotation than others, while negative documents were 31% less likely. This is intuitively consistent with the results for *thumbs up* annotations, although the effect size is smaller. Interestingly, this *thumbs down*-positive document relationship is also observed for the Leaders' Debate, where positive documents were 30% more likely to receive *thumbs down* feedback ($p < 0.05$). We also observed significant associations for neutral X Factor documents, which were 28% more likely to receive a *thumbs down* ($p < 0.05$).

For X Factor feedback, *thumbs up* feedback was significantly less than expected where participants were (i) aged 25 or older and (ii) male ($p < 0.001$). We observed the inverse for *thumbs down* and found this pattern to be most prominent in positive documents. An interesting observation from the debate is that those who were familiar with microblog search were 80% more likely to *thumbs up* a neutral document ($p < 0.05$), perhaps due to a higher level of trust placed in informative documents.

Across both topics, we see positive documents are negatively received by participants and negative documents are positively received, reinforcing what we see at algorithm level. Many of the documents classified as positive are those where the sentiment is explicit, often a few words stating support for a topic entity, offering little in the way of content. We observed that humorous and critical content tends to be negative; perhaps this was valued highly by participants.

Participant Sentiment

We observed no significant between-subjects effect for any of the prior participant sentiment categories with respect to different sentiment filtering algorithms for either overall stream feedback or T_{Net}. For the Leaders' Debate, positive participants were less likely to give *thumbs down or thumbs up* feedback for either positive ($p < 0.05$) or negative ($p < 0.001$) documents. Indeed, positive participants were more than three times less likely to *thumbs down* a negative document and only half as likely to *thumbs down* a positive document. Neutral participants on the other hand were more than 50% more likely to *thumbs up* a document regardless of sentiment.

The effect sizes observed for the X Factor were smaller, though in this case we saw a higher likelihood of positive participants annotating positive documents as *thumbs up*, and a lower likelihood of positive participants annotating positive documents as *thumbs down* ($p < 0.001$). Negative participants were less likely to *thumbs up* a positive document ($p < 0.001$) though other effects related to negative participants were small, or not significant.

The patterns for the X Factor and the Leaders Debate are quite different, although positive documents are consistently perceived the worst in each grouping. The larger effect appears to be a difference in the way participants of different prior sentiment approach the task, rather any effect related to content sentiment.

Conclusion

We have described a system and methodology for examining sentiment in real-time microblog search scenarios. We took two topics, a political debate and an entertainment television show, and conducted a series of laboratory user studies. The largest effect we observed was for positive content which consistently receives negative feedback. This is perhaps to do with participant dissatisfaction with positive content, or perhaps the absence of other types of sentiment. Our `neg` algorithm performs similarly to the `control` stream despite having much less neutral and positive content. We speculate that negative content is valuable to the searcher. The `control` proves to be a strong baseline.

We found that comparing feedback to participant prior sentiment reveals significant patterns. We conclude also however, that these profile variables have a stronger association with task feedback in general, rather than any sentiment-specific aspect.

Although the effects we see are mixed, it is clear that sentiment in a real-time search system is a measurable quantity, that we can use sentiment to produce significant reponses from users, and that we can capture this response effectively with our experimental methods. This is a promising result for automated sentiment analysis in real-time microblog search.

Acknowledgments

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

References

Belkin, N., and Croft, W. 1992. Information filtering and information retrieval: two sides of the same coin? *Communications of the ACM* 35(12):29–38.

Bermingham, A., and Smeaton, A. F. 2011. On using Twitter to monitor political sentiment and predict election results. In *SAAIP - Sentiment Analysis where AI meets Psychology workshop at the International Joint Conference on Natural Language Processing (IJCNLP) November 13, 2011, Chiang Mai, Thailand*. in press.

Broder, A. 2002. A taxonomy of web search. In *ACM Sigir forum*, volume 36, 3–10. ACM.

Churchill, A.; Liodakis, E.; and Ye, S. 2010. Twitter relevance filtering via joint bayes classifiers from user clustering. CS 229 Final Project, Stanford University.

Deller, R. A. 2011. Twittering on: Audience research and participation using twitter. *Time* (2009):216–245.

Efron, M., and Golovchinksy, G. 2011. Estimation methods for ranking recent information. In *SIGIR '10: Proceedings of the 34th Annual ACM Conference on Research and Development in Information Retrieval*. New York, NY, USA: ACM.

Efron, M. 2011. Information search and retrieval in microblogs. *Journal of the American Society for Information Science and Technology*.

Macdonald, C.; Santos, R. L. T.; Ounis, I.; and Soboroff, I. 2010. Blog Track research at TREC. *Sigir Forum* 44.

Massoudi, K.; Tsagkias, M.; de Rijke, M.; and Weerkamp, W. 2011. Incorporating query expansion and quality indicators in searching microblog posts. *Advances in Information Retrieval* 362–367.

Mishne, G., and de Rijke, M. 2006. A study of blog search. *Advances in Information Retrieval* 289–301.

Mishne, G. A. 2007. *Applied Text Analytics for Blogs*. Ph.D. Dissertation, University of Amsterdam, Amsterdam.

Ounis, I.; Macdonald, C.; and Soboroff, I. 2008. Overview of the TREC-2008 Blog Track. In *Proceedings of the 17th Text REtrieval Conference (TREC 2008)*.

Seki, Y.; Ku, L.-W.; Sun, L.; Chen, H.-H.; and Kando, N. 2010. Overview of multilingual opinion analysis task at NTCIR-8. 209–220.

Sriram, B.; Fuhry, D.; Demir, E.; Ferhatosmanoglu, H.; and Demirbas, M. 2010. Short text classification in Twitter to improve information filtering. In *Proceeding of the 33rd international ACM SIGIR conference on research and development in information retrieval*, 841–842. ACM.

Teevan, J.; Ramage, D.; and Morris, M. 2011. # TwitterSearch: a comparison of microblog search and web search. In *Proceedings of the fourth ACM international conference on Web search and data mining*, 35–44. ACM.