Sentiment Analysis and Real-time Microblog Search

Adam Bermingham

A dissertation submitted in fulfilment of the requirements for the award of

Doctor of Philosophy (Ph.D.)

to the

Dublin City University
School of Computing

Supervisor: Prof. Alan F. Smeaton

21st December, 2011

Declaration

I hereby certify that this material, which I now submit for assessment on the programme of study leading to the award of Doctor of Philosophy is entirely my own work, that I have exercised reasonable care to ensure that the work is original, and does not to the best of my knowledge breach any law of copyright, and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the text of my work.

Signed:

ID No:

Date:

Abstract

This thesis sets out to examine the role played by sentiment in real-time microblog search. The recent prominence of the real-time web is proving both challenging and disruptive for a number of areas of research, notably information retrieval and web data mining. 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 query at a given point in time, automated methods are required to enable users to sift through this information. As an area of research reaching maturity, sentiment analysis offers a promising direction for modelling the text content in microblog streams.

In this thesis we review the real-time web as a new area of focus for sentiment analysis, with a specific focus on microblogging. We propose a system and method for evaluating the effect of sentiment on perceived search quality in real-time microblog search scenarios. Initially we provide an evaluation of sentiment analysis using supervised learning for classifying the short, informal content in microblog posts. We then evaluate our sentiment-based filtering system for microblog search in a user study with simulated real-time scenarios. Lastly, we conduct real-time user studies for the live broadcast of the popular television programme, the X Factor, and for the Leaders Debate during the Irish General Election. We find that we are able to satisfactorily classify positive, negative and neutral sentiment in microblog posts. We also find a significant role played by sentiment in many microblog search scenarios, observing some 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.

Acknowledgements

First, I'd like to thanks my supervisor, Alan, for his guidance throughout my time as a PhD student. He gave me encouragement to identify and pursue my own research questions and goals, and to question existing methodologies. He has taught me the value of human factors in technological research and development, something I anticipate will stay with me throughout my career. I am also grateful for Alan's assistance in organising the logitistics of the user studies, not least offering his own home for what he billed as "The CLARITY X Factor Tweetfest". Last, but not least, he has educated me (repeatedly) in the difference between "effect" and "affect", and I assure him I have finally got the hang of it.

Thanks to the administration staff in CLARITY DCU, Ann Marie and Deirde, for helping me organise my user studies. Organising studies around real-time events is tricky at the best of times, and there were some hairy moments when we had to postpone one study on account of snow. Thanks also to my annotators, and all those who participated in the user studies. I appreciate especially the volunteers who had negative prior sentiment towards the event topics.

I would also like to thank my proofreaders: Caroline, Dave, Ronan and Breffní. Your feedback was very helpful and you'll be happy to know I've now nailed the spellings of *corellation*, *suspition*, *diffrentiate* and *occurance*. Since your feedback, I also, in writing this thesis, have tried, as best I can, to include more commas, wherever they are necessary.

CLARITY and CDVP is a great place to work, and there are many who have made my experience during my PhD an enjoyable and educational one. I would like to single out James, Pete, Cathal, Daragh and Colum, who each at various stages provided advice which I found very valuable in my research. A special thanks goes to Edel O'Connor, who has had to put up with me sitting beside her for four years. Edel has always been there to discuss any problems I was having with my research, and has provided top-notch company for the long days in the lab. I would also like to thank others in the lab that have made the PhD experience enjoyable: Niamh, Kevin, Neil, Paul, Phil, Aiden, Eamonn, Ham, Eoin... and the rest, you know who you are!

I would like to thank my parents, Anne Marie and Anthony, for the support they have given me, not just during my PhD, but down through the years. I appreciate the genuine interest, support and encouragement they have shown, and I can safely say they have provide me with a constant inspiration.

Lastly, and most importantly, I'd like to thank my wife, Susan. Throughout my PhD, Susan has offered me counsel, support, food, coffee, and in general has displayed an admirable level of understanding and tolerance of mood swings, stress and 24/7 work schedules! Thanks Susan, I couldn't have done it without you.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Over the last 10 years, user-generated content has come to dominate a large portion of the web (Wunsch-Vincent and Vickery, 2007). 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. However, the information in this, social web, is unlike much of the information in the traditional web. One of the primary differences is that social information has a characteristically high degree of subjectivity. This has inspired research in the area of automated *sentiment analysis*: methods for automated detection of negative and positive emotions, opinions and other evaluations in text.

In this research we are focused on the *real-time web*. This 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 (i.e. offline world) either at, or close to that time. In terms of user-generated content, this 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 scheduled times, or can occur spontaneously, so too does user-generated content have a prominent time component. Examples of scheduled real-world events would be sporting contests and television programmes. Examples of spontanous real-world events would be breaking news stories, disasters and civil disturbances.

The microblogging service, Twitter[1], is a good example of information comprising the real-time web. Twitter allows users to publish short text documents, or "posts", which appear in their followers' feeds, and may appear in search results. Twitter users write about a wide variety of topics including both scheduled and spontaneous real-world, real-time events. The diversity of content, and the abundance and availability of data, mean that Twitter provides us with a unique opportunity to analyse sentiment in real-time, in a way not before possible. Throughout this thesis we use Twitter as a case study for sentiment in the real-time web. In applying sentiment analysis to the real-time web, and in specific microblog 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 meaningful information about the world around us.

In the years since the dawn of the Internet, information access systems have been at the core of the user experience. They have applied order to the web, and empowered the Internet user to navigate it effectively to satisfy their information needs. Recently, real-time social content is more and more becoming part of our perception of the real world. Yet, it is unclear how to develop systems to best enable users to explore this information. Also, the role played by subjectivity in real-time information systems is largely unknown. In this thesis we explore the potential to harness this subjective power using automated sentiment analysis to allow a user to understand the social web in real-time.

Our experiments apply sentiment analysis in a real-time microblog search system. This is not something that has been possible until now in any meaningful way, and so we are presented with a unique and novel avenue for research.

1.1 Motivation

Social computing has become pervasive in our society. At the time of writing, the popular social networking site, Facebook[2], has over 600 million active users (Carlson, 2011); Twitter has approximately 200 million registered users (Baird, 2011). The day-to-day management of an online persona and consumption of information from social sources

[1] `http://www.twitter.com`
[2] `http://www.facebook.com`

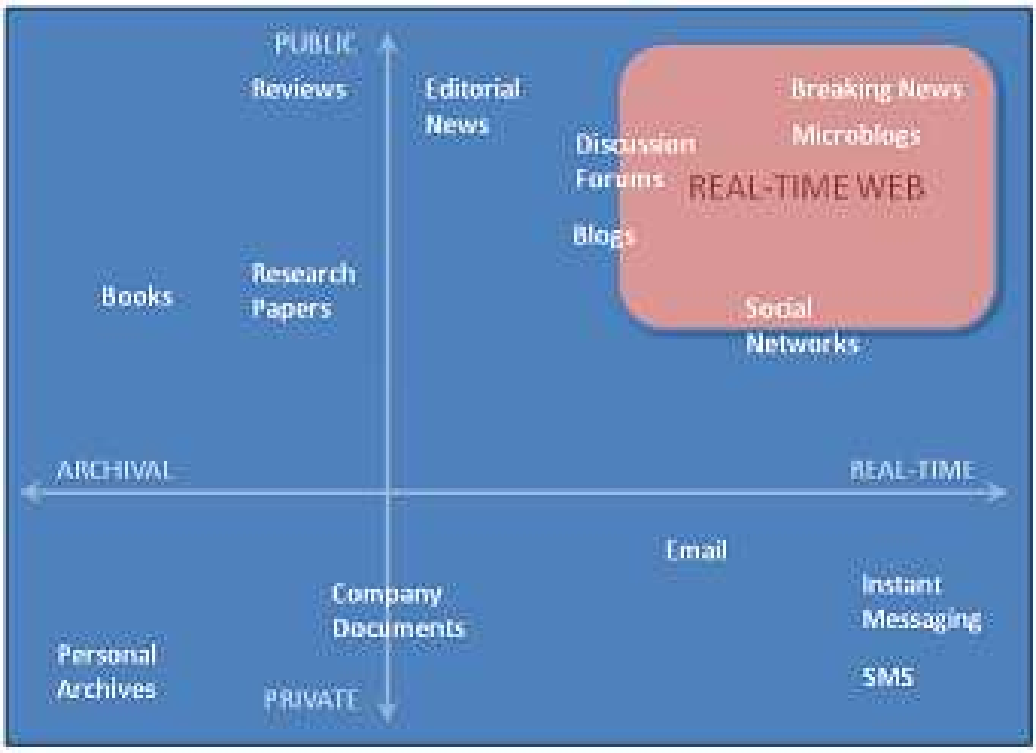


Figure 1.1: A conceptualization of the real-time web in terms of digital content

have become commonplace.

The recent growth in the volume of data in the real-time web, specifically on Twitter, is staggering. At least one website has recently measured the rate of Twitter posts, or *tweets*, being published as 2 billion per month, or 64 million per day, and increasing (Pingdom, 2010). The improvements in smartphone and tablet technology, combined with affordable pricing, mean that the barriers to access of the social web have been considerably eroded. User-generated content can now be created and consumed instantaneously, wherever the user is. For example, if a user has a thought about a product they are using or has captured an interesting photograph concerning a breaking news story, they can instantly upload this to the web for others to see. Similarly, if a user has an information need associated with an unfolding event, they can find relevant commentary on Twitter, moments after it has been authored.

But what portion of this deluge is relevant to a given topic interest? During the recent 2010 FIFA World Cup in South Africa, even the early-stage matches saw activity in the region of hundreds of thousands of tweets per match. Similar activity was seen during

the NBA play-offs of the same year. High levels of activity are also seen in relation to unfolding news stories, and live television. Clearly users need to be assisted in their search for relevant information — users are presented with an *information overload problem*. Given a user's real-time information need, and the abundance of real-time information, how can we sample this stream to the benefit of the user?

For some time there have been methods of near-instantaneous computer-mediated communication (CMC). 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. See Figure 1.1 for a conceptualisation of digital content in the real-time web.

Automated sentiment analysis as a fundamental technique is reaching an age of maturity. There are now established methodologies, in particular for machine learning techniques, for obtaining accuracies comparable with the traditionally easier task of topical classification. Now that the research community understands many of sentiment analysis technology's capabilities and limitations, we endeavour to demonstrate its benefit in application areas. However, even leading web search provider, Google[3], conceded recently that they have had trouble demonstrating an improvement in web search performance using sentiment analysis:

> *"So far we have not found an effective way to significantly improve search using sentiment analysis. Of course, we will continue trying."* (Singhal, 2010)

For the task of search of 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. The social commentary either tends to be partisan (e.g. politics, sports) or critical/reviewing

[3] `http://www.google.com`

(e.g. television). Perhaps in the real-time social web, users' real-time needs have a prominent sentiment component. This thesis work describes and evaluates a system for allowing users to view a stream of real-time social content from the microblogging site, Twitter, while observing events. Our evaluation goal is to better understand the role that sentiment plays in such an information access system.

Search on Twitter[4] is dominated by inverse-chronologically ordered results, filtered by keyword. In this model, the assumption is that recency is the single most discriminating factor between relevant documents. In the case where relevant documents are being authored at a great rate, there will be many more before a user even has time to finish reading the search results. This simple model does not scale well, and real-time microblog search is still an unsolved problem. Perhaps real-time streams of user-generated content are destined to be passively observed, or perhaps a more active search interaction is preferable. The problem definition and methodologies are still in flux. By enriching the documents with sentiment information, there exists an opportunity to employ more sophisticated methods to help users find useful information, for example by ensuring a level of diversity and representativeness of sentiment in the results list, or choosing content which aligns with the user's own personal sentiment.

This research comes at the convergence of a number of developing technologies: the social web, ubiquitous computing, real-time information retrieval and sentiment analysis. It is the intersection between these technologies, the abundance and availability of data and the dearth of research into sentiment-based strategies for real-time information access that motivate this thesis research.

1.2 Hypotheses and Research Questions

There are two objectives at the core of this thesis:

1. We aim to demonstrate how a real-time sentiment-based information access system can be built and evaluated methodically. This includes development of the fundamental sentiment analysis, as well as development of a rigorous, real-time, user-based evaluation methodology.

[4] `http://search.twitter.com`

2. We aim to explore the relationship that sentiment has with users in a real-time context, drawing conclusions about users' profiles and preferences, and assessing the successes and failures of our system.

More formally, we state our research focuses as hypotheses and research questions in the following sections.

1.2.1 Hypotheses

1. *Using a sentiment-based sampling strategy to create a stream will elicit significantly different responses from users to a random sampling method in a real-time event scenario.*

The first, and primary hypothesis, concerns the effects associated with the percieved utility of sentiment in a document stream, in the context of a live event. The key assumption in the first hypothesis is that certain types of sentiment will be of interest to the information seeker; others will not. In order to evaluate this hypothesis we need a robust underlying framework for analysing sentiment in microblog posts. This leads us to our second hypothesis:

2. *The succinctness of microblogs allow us to efficiently mine their sentiment, despite their short length and informal nature, using supervised learning approaches.*

The second hypothesis addresses the domain-specific challenges involved in this area of research. We assume that users, when forced to be brief, are concise in their language, thus providing us with information-dense, explicit text documents. Following from this assumption, we suppose that statistical methods, such as supervised learning, may be used effectively to mine the sentiment contained in microblog documents, even though the shortness of the documents presents us with a potential problem, due to presence of only a small number of features in any given document.

1.2.2 Research Questions

In order to verify these hypotheses, we must evaluate a number of important research questions:

1. *In what ways do the natural language and the textual conventions used in microblog text differ from that used in other types of user-generated content?*

2. *What effect does the nature of microblogs have on sentiment analysis using supervised learning for microblog posts, compared to traditional, longer document classification? What comprises an optimum feature set and classification strategy?*

3. *How do we model real-time microblogging as an information access system? How may this be most effectively combined with the classification strategy established in (2)?*

4. *How may the system proposed in (3) be evaluated with respect to users' real-time information needs? Do sentiment-based algorithms differ significantly from a baseline sampling approach?*

5. *Do users' demographics and preferences significantly affect their perception of sentiment? Which types of sentiment have the most profound impact?*

6. *Is sentiment a predictor of whether individual documents will be regarded as important by users?*

1.2.3 Research Contributions

The first contribution of this research is a thorough review of literature concerning real-time information access systems, with particular focus on the social web. We also review the state-of-the-art in research concerning microblogs, and sentiment analysis on short and informal text.

Our second contribution is a system and model for integrating sentiment into a real-time, event-based microblog stream. This includes a methodology for creating high-quality training data, and a rigorous evaluation of applied machine learning techniques for performing sentiment analysis in this context, drawing comparison with other domains.

Thirdly, we propose and perform a method for real-time system evaluation using real users in laboratory settings. We offer a number of evaluation metrics and provide an evaluation of the role of sentiment with respect to (i) users, (ii) stream sampling algorithms, and (iii) document features.

1.2.4 Thesis Structure

The structure of this thesis is as follows:

- **Chapter 1**: In this, the current chapter, we introduce the concepts of sentiment analysis, the real-time web and microblogging, offering motivation and justification for our work. We present our research aims, hypotheses and research questions.

- **Chapter 2**: This chapter contains our survey of related work and presents a high-level overview of our research, introducing our experimental methodology. Our methods use real-time user feedback to establish the quality of content in the stream and our experiments are structured so that we take an incremental approach towards answering our research questions.

- **Chapter 3**: In this chapter we present the design and architecture of our experiment system, *Channel S*. We detail the specification and implementation of the system, and describe how it supports our evaluation methodology.

- **Chapter 4**: This chapter specifically concerns sentiment analysis using supervised learning for microblog posts. We survey the related work in the area and present our experiments and findings, comparing and contrasting with data from three other domains. We also describe the materials and methods we develop to construct our body of training data for our experiments.

- **Chapter 5**: In this chapter we describe our user study for evaluation of sentiment in simulated real-time search scenarios. We use labelled sentiment data from the previous chapter's experiment to control the sentiment in the search tasks. We present findings and discussion from experimental feedback, noting a number of significant sentiment-related patterns concerning topics, users, streams and documents.

- **Chapter 6**: In our final experimental chapter, we describe our live, event-based laboratory user studies. This experiment integrates the automated sentiment analysis into the search system, deploying the system in real-time during (i) two broadcasts of the television show, the X Factor (series 7, 2011), and (ii) the Leaders' Debate during the Irish General Election, 2011. We present and discuss our findings in each

of our studies, comparing and contrasting our observations for each of the events throughout. In this chapter, we also describe the *GE11 Twitter Tracker*, a live system we developed for an Irish news website that allowed users to monitoring public political sentiment on Twitter in real-time during the Irish General Election.

- **Chapter 7**: In our final chapter we summarise our conclusions with respect to our hypotheses and research questions. We also reflect on the work as whole and present directions for future work.

- **Appendices**: In our appendices we present our research materials such as topics, annotation guidelines, ethics materials and user surveys. We also summarise our published work which has served as a precursor to this research, as well as our published research which has directly contributed.

The chapters are structured to accomodate readers with different levels of interest and expertise. Those uninterested in machine learning may wish to skip Chapter 4; those who wish to get a high-level overview may wish to simply read Chapters 1, 2 and 7; non-technical readers may wish to skip Chapter 3. For those solely interesting in supervised sentiment classification it is recommended to read Chapter 4 as well as the sentiment configuration for our real-time studies (Section 6.2.4).

Chapter 2

Related Work and Thesis Overview

In this chapter, we give an overview of sentiment analysis in the real-time web and in particular, microblog search. This is a relatively new area of research and, in order to establish a solid foundation on which to evaluate our thesis, we look at related information retrieval research. Specifically, we model microblog search as an information filtering task and propose an evaluation methodology based on established methods for interface evaluation for information retrieval systems, allowing us to perform experiments with sentiment as a controlled variable.

We begin in the following section with a review of information retrieval and related work and formulate our microblog search problem. Then, in Section 2.2, we give an overview of sentiment analysis, with a focus on sentiment-oriented information systems. We discuss our evaluation methodology in Section 2.3 and conclude in Section 2.4.

2.1 Information Retrieval

> *"Information retrieval is a field concerned with the structure, analysis, organization, storage, searching, and retrieval of information."* (Salton, 1968)

Although this definition of information retrieval dates from the early days of the field, it is still applicable to modern information retrieval in the context of the web and social

search. A recent textbook says of Salton's definition:

> *"Despite the huge advances in the understanding and technology of search in the past 40 years, this definition is still appropriate and accurate."* (Croft et al., 2009)

Croft *et al.* continue to add that modern information retrieval concerns the tasks of *question answering*, *filtering*, *ad hoc search* and *classification*, among others. In this section, we explore microblog search as an information retrieval problem. We review related information retrieval research from the area of information filtering as well as more recent work specifically concerning microblog content.

2.1.1 Information Needs in Social Content

In information retrieval systems, users typically expect the system to provide them with documents they will find useful given their information need. More formally this has been described as the resolution of an *Anomolous State of Knowledge* (ASK) (Belkin and Croft, 1987). In textual systems, a user's information need is approximated by a short query string which the system uses to suggest relevant documents. The most familiar contemporary example of this is web search, a task completed regularly by Internet users.

Much newer and less well understood than web search, is the task of microblog search. Microblog search is perhaps most closely related to the more mature field of blog search, which has been a focus of search in user-generated content in recent years. Context and motivation for blog search comes from analysis of the information needs in blog search query logs. Mishne and de Rijke (2006b) found that the two primary categories of blog search query are *concept* and *context* queries. Whereas *concept* queries concern a topic or area of interest (e.g. "growing food in small spaces","sports cars"), *context* queries aim to find commentary on real-world entities such as products or public figures (e.g. "the oscars", "barrack obama"). They remark how this significantly differs from web search information needs, which are described as *informational*, *navigational*, or *transactional* (Broder, 2002). This taxonomy for web search queries is still considered standard today.

The prevalance of *context* queries in blog search has inspired much work on opinion-based search. With *context* queries, the information need is described as the wish of the

user to find subjective commentary about an (often topical) real world entity of interest. The observation that the content in blogs is often subjective, has led to efforts to formulate this information need as one specifically seeking subjective commentary about the entity in question.

As we will see in Chapter 4, popular microblog topics largely conform to the notion of *context* queries, and thus it is our intuition that a similar desire for opinionated commentary to that exhibited in blogs is prevalent in microblogs. We cover opinion-based search in more detail in Section 2.2, but first let us consider the specific problem of microblog search.

2.1.2 Problem Description

There are a number of similarities between search on web and blogs, and microblog search. The units of information are similar — discrete documents. The queries are also similar — short text statements of information need. There are however also differences; microblog documents have a length constraint (just 140 characters) and, as a real-time communication platform, have a strong temporal component.

There are two types of microblog search query we might consider:

- *Ad-hoc:* A user has an instantaneous information need, at a specified point in time, and desires a single set of documents.
- *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. At the time of writing, such an effort is underway as groups prepare to participate in the Text REtrieval Conference (TREC) Microblog Track[1].

It is, however, the latter of these types of query, *persistent* queries, which are of interest to us. These queries allow people to track live events such as television programmes, breaking news stories, debates, sports and many other types of real-time event, as the event is unfolding. A common form of this is using a computer or mobile device to follow

[1] `https://sites.google.com/site/trecmicroblogtrack/`

an event on the social web, while also watching the event on television. This practice is known as *second-screen viewing*. This new type of real-time social information access augments the viewer's experience of a live event. Developing systems which effectively enable users to engage with such real-time social content is an exciting new challenge for information retrieval research.

Persistent microblog queries largely fall into the category of context queries as defined for blog search. 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.

Thus, we may state the problem:

> *Given a stream of microblog documents, S, how do we create a derivative stream S' which consists of documents from S and which optimally matches the user's stated information need.*

Although we can summarise the problem succinctly, it is a complex task. Interaction variables around the user scenario deserve consideration, as do methods for determining the perceived quality of a user's search streams. It is also unclear whether a one-size-fits-all general solution is appropriate, or whether users have radically disparate preferences.

2.1.3 Information Retrieval and Information Filtering

Web search, an ad-hoc search task, has been at the core of the web experience since the dawn of the Internet. As another subtask of information retrieval, information filtering has also had a role to play, although it has been somewhat overshadowed by ad hoc search. Information filtering is generally concerned with removing non-relevant documents in a stream of documents for a user, rather than actively searching for documents or information, as is the case in ad hoc search. A seminal paper which addresses the distinction between ad hoc search (referred to simply as IR) and information filtering concludes:

> *"...most of the issues which appear at first to be unique to information filtering, are really specializations of IR problems."* (Belkin and Croft, 1992)

Throughout they argue that information filtering and ad hoc search are in essence very similar tasks. This is important for us, as useful approaches from ad hoc search literature may potentially be effectively adapted to our microblog filtering problem.

In the same paper, Belkin and Croft define an information filtering system with respect to six criteria. Let us consider each of these and how they relate to microblog search:

1. An information filtering system is an information system designed for unstructured or semi-structured data.
2. Information filtering systems deal primarily with textual information.
3. Filtering systems involve large amounts of data.
4. Filtering applications typically involve streams of incoming data.
5. Filtering is based on descriptions of individual or group information preferences, often called profiles. Such profiles typically represent long-term interests.
6. Filtering is often meant to imply the removal of data from an incoming stream, rather than finding data in that stream.

Microblog searches and documents are solely textual and contain natural language content, satisfying criteria (1) and (2). With upwards of several hundred tweets per second on average (Garrett, 2010), Twitter as a microblogging service comfortably satisfies criterion (3). Microblogging's instantaneous publishing and established stream-style interaction patterns conform to criterion (4). The abundance of data and immediacy of information needs mean that undesirable data must be omitted from the stream to be monitored satisfying criterion (6).

Criterion (5) concerns the reasoning used to include (or exclude) documents from the stream. Traditionally this is thought of in information filtering literature as a profile which over time can be learned through feedback, providing a personalised stream. In this work we use sentiment-based filtering criteria so that we may isolate and evaluate the perceived effect of sentiment in the stream. Although different, this task still conforms to classical information filtering and we may consider it as such.

Early research formulated the information filtering problem as "*document routing*" where the goal was to determine the relevance of documents to topics given some training relevance judgments and topics and relevance feedback at subsequent time intervals (Schütze et al., 1995). The primary difference with ad-hoc search, is that assessments of relevance must be made temporally (and hence sequentially), rather than at a set level, or by ranking. Schütze *et al.* note that the task is essentially a document classifcation problem for binary relevance. Harman (1995) provides another description:

> *"In the routing task it is assumed that the same questions are always being asked, but that new data is being searched."*

The applications which motivated research into document filtering were personalised news services and identifying new and relevant literature.

Research on filtering moved from document routing, to the more difficult problem of adaptive filtering. In adaptive filtering, few or no document judgments are known at the start, and the focus is on leveraging the information contained in online relevance feedback to construct a relevance profile. It was found experimentally that adaptive filtering systems could perform as well as previous routing or batch filtering approaches, despite requiring considerably less training data, as shown by (Robertson and Hull, 2000). This research challenge evolved into novelty detection where the goal was to find new ("novel") relevant information in a temporal stream, see for example (Allan et al., 2003; Gaughan and Smeaton, 2005; Yang et al., 2002). Clarke et al. (2008) present some of the considerable challenges in evaluating systems with respect to novelty in results (and the related concepts of diversity and redundancy), a problem which has arguably slowed progress in this area.

One of the most salient example applications of filtering systems, is that of spam-filtering, an area more broadly referred to as *adversarial information retrieval*. Cormack (2007) provides a review of the area. Notable approaches include Naive Bayes classification (Androutsopoulos et al., 2000), case-based reasoning (Cunningham et al., 2003) and employing support vector machine and maximum entropy model classification (Zhang et al., 2004).

The above approaches may all be thought of a *content-based filtering systems*. A related approach to a similar problem, is *collaborative filtering*, or *recommender systems* as they are now more commonly known. In these systems the preferences of similar users are used to identify documents of potential interest. A common example is an e-commerce site suggesting an item to a user based on what other users have purchased who have a similar buying history. See Resnick et al. (1994) for an example of early collaborative filtering work. More recent surveys of the literature can be found in Su and Khoshgoftaar (2009) and Adomavicius and Tuzhilin (2005).

While recommender systems have enjoyed much success in recent years, in this research we evaluate sentiment as a content-based filtering mechanism. We use the filtering metaphor to assess sentiment's role in real-time microblog access. As we will see, the architecture and experimental set up closely resemble that of an information filtering task. Just like filtering tasks, whether relevance or spam filtering, we use supervised learning to remove messages from the stream according to sentiment profiles. As we will see, this also allows us to evaluate other aspects of the system too, not simply the filtering algorithms. Although collaborative systems are not a focus of this work, collaborative sentiment-based filtering is a potential long-term research avenue.

2.1.4 Search Tasks on Microblogs

At this stage we have reviewed related information retrieval literature. Now let us examine some recent research in the area which specifically addresses the task of microblog search. As a new area of focus for the information retrieval community, how best to tackle microblog search is very much an open research question. Indeed, most of the research we discuss here is from the previous 18 months at the time of writing.

One significant work which has tackled microblog search is Massoudi et al. (2011). Massoudi *et al.* use query expansion and quality indicators to extend a language model information retrieval approach to microblogs. Their quality indicators build on previous work which demonstrated the benefit of using credibility indicators in blog search (Weerkamp and de Rijke, 2008). They find that both query expansion and quality metrics improve retrieval performance over a Boolean search recency-ranked baseline.

Another notable work used 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 ad hoc (i.e. instantaneous) and derive their methodology from traditional static information retrieval evaluation.

Teevan et al. (2011) provide a valuable comparison of web search and microblog search through query log analysis. They note that Twitter queries tend to be shorter than web queries and are likely to be related to hashtags. Hashtags are terms in microblog posts preceded with a hash character ("#"), used by the author to add a keyword or tag to the post. They also remark on the prevalence of questions in tweets, with 17% of the tweets in their corpus containing a question mark. A finding which bears particular relevance to our work is the following:

> *"Twitter search is used to monitor content, while Web search is used to develop and learn about a topic."*

This reinforces our assertions about the importance of monitoring, or the persistent nature of some queries. This is a usage pattern which some of the aforementioned, traditional-style ad hoc evaluations struggle to address.

Some recent works tackle microblog search as a filtering problem. Sriram et al. (2010) filter tweets using a Naive Bayes classifier to categorise tweets into general categories such as "news" and "events". Churchill et al. (2010) use social information to perform user clustering and generate individual user profiles. This area is however largely unexplored, perhaps due to the poorly understood information needs of persistent queries, and difficulties in evaluating such.

An emerging task is the detection of the important themes in a set of microblog documents. This has been approached as a clustering problem (O'Connor et al., 2010b) and as a topic retrieval problem, where the objects of retrieval are hashtags (Efron, 2010). Similar work considers the aggregation of a stream of microblog posts during an event as

a summarization problem, where the end goal is a filtered stream of tweets (Takamura et al., 2011). It should be noted that this summarization is only employed retrospectively, and as such does not make provisions for real-time use cases. Examples of more general knowledge discovery tasks in microblogs include personalised ranking of news feeds using Twitter (Phelan et al., 2011), recommending people to follow (Hannon et al., 2011) and earthquake event detection (Sakaki et al., 2010).

In summary, we see there are several tasks in microblog search which deserve attention: ad hoc search, persistent search, question answering, topic extraction, summarisation and discovery of other real-world information. Each of these areas is a new and challenging area of research with open research questions. Our task, persistent search, is perhaps one of the least-well understood; there is a lack of cohesion in how this research problem is formulated and how approaches are evaluated experimentally.

2.2 Sentiment Analysis

The central focus of this thesis research is to investigate the role that sentiment plays in real-time microblog search scenarios, with particular focus on the task of persistent search. Until now, sentiment has been somewhat of an unknown quantity in terms of information systems. It is tempting to think that sentiment, a highly subjective notion, is a characteristic which users might find discriminative when it comes to their information preferences. In this section, we digress from the task of microblog search to consider related work on sentiment and the nature of sentiment as it persists in data. For background related specifically to supervised learning for sentiment analysis, see Chapter 4.

2.2.1 Background

Sentiment analysis suffers somewhat from lack of convergence in terminology. In this and the following section, we give an overview of the terminology, and historical background to sentiment analysis, as well as the tasks and problems in the area, drawing on the overview of the area given in Pang and Lee (2008).

Subjectivity analysis is perhaps the most broad term used to describe the general area of identifying subjective, opinionated or emotional content in text. It can be extended

to include such notions as evaluation and speculation i.e. an appraisal of an entity's performance or value, or expectations of its performance or value at a future point in time. There are however three notions that comprise subjectivity analysis: *opinion*, *sentiment* and *subjectivity*. Opinion concerns an opinion expressed in text, often consisting of a target, or target feature, associated with a given opinion and an opinion holder. This idea is frequently used, for example, in the mining of product reviews.

Subjectivity in textual content, is content which is distinct from objective fact in that it communicates the private states of the author. Although evaluation, emotions and speculations can be included under this umbrella, research in this area is largely concerned with the identification or extraction of opinion-oriented language in text.

Lastly, sentiment itself is used most frequently when referencing the *valence* (or *polarity*) of content towards a given topic, i.e. positivity or negativity. In sentiment, often the focus is more on an evaluative perspective on the topic at hand as the author wishes to convey how favourably they consider the topic. Sentiment analysis has been used to narrowly define the area of subjective research concerned with this evaluative text, particularly using review data. It is now becoming more and more common to use sentiment analysis to refer to the broader task of computationally identifying opinion, subjectivity and sentiment in text.

These popular formulations of the sentiment analysis and subjectivity analysis problem appear to have been around since 2001. Previous to this, much of the work was in the are of distinguishing subjective and objective content in a given narrative. Important work in this comes from Weibe (Wiebe, 1990, 1994; Wiebe and Bruce, 2001). At this time the task was as much about tracking narrative viewpoints as isolating the factual content in text. This predates (i) significant modern advances in machine learning technology and other statistical techniques, (ii) the explosion in textual data available in the World Wide Web and ultimately (iii) the commercial demand for monitoring, managing, analysing and understanding this data.

Prior to this Ekman had begun formulating what is now known as the Ekman's Basic Emotions. These are: *anger*, *fear*, *sadness*, *enjoyment*, *disgust*, and *surprise*. Ekman devised this classification of emotions based on cross-cultural analysis of facial expressions and argued that these emotions are intrinsic to all humans and not culturally derived. An

overview of this work can be found in Ekman (1989). He states that:

> *"These findings forced me to reject my previous beliefs that: (1) a pleasant-unpleasant scale was sufficient to capture the differences among emotions; and (2) the relationship between a facial configuration and what it signifies is socially learned and culturally viable"* (Ekman, 1992)

This is a cautionary warning for approaches which choose to use a pleasant-unpleasant (or positive-negative) scale to model human sentiment, in that it may not have the ability to model the complexity of human emotions. This however must be balanced with the fact that such emotion is difficult to measure accurately in text and more naive, but simpler models, can yield promising performance. Some recent systems user emotion taxonomies to model sentiment. Cambria et al. (2010) propose a resource for building emotional context into WordNet; Aman and Szpakowicz (2007) report positive results in annotator agreement for labelling Ekman's emotional states in text; Bollenet al. use another emotion system, known as the the Google Profile of Mood States (GPOMS) to model public mood and predict the stock market (Bollen et al., 2009, 2011). These mood states are: *calm*, *alert*, *sure*, *vital*, *kind*, and *happy*. They used this in conjunction with a polarity-based tool, *OpinionFinder*.

2.2.2 Tasks

At this stage it is useful to give a brief overview of the tasks and applications that fall under the umbrella of sentiment and subjectivity analysis:

- *Subjectivity identification:* Identifying subjective text in order to distinguish it from objective/factual content
- *Polarity classification:* When subjectivity is assumed, classifying content as one of positive or negative, or assigning an ordinal label to content on a positive-negative graded scale, or similar
- *Emotion recognition:* Identifying distinct human emotions in textual content, beyond binary notions of positivity and negativity or subjectivity and objectivity

- *Opinion extraction:* Extracting the opinion itself, often as a tuple containing the opinion holder, opinion valence (and possibly strength), and opinion target
- *Joint topic-sentiment analysis:* The relevance of text to a topic is unknown so the task involeves both topical relevance modelling, as well as for example, subjectivity identification or polarity classification

These tasks are often modelled as the following problems:

- *Summarization:* Abbreviating the textual content, either through abstractive, extractive or visual means, to succinctly display subjective content
- *Extraction:* Extracting information from the content
- *Retrieval:* Ranking documents in response to some sentiment-oriented query, for example topic-opinion search
- *Classification:* Automatically labeling documents as, for example, positive or negative, or subjective or objective, often using machine learning techniques, either unsupervised or supervised
- *Measurement:* Using one or more of the aforementioned approaches, derive metrics for quantifying sentiment so that sentiment maybe be monitored, measured and used in statistical models, for example, for predicting other data series

We discuss some of the tasks and related work in more detail in the following section, as well as the in the background section of Chapter 4.

2.2.3 Related Work

Sentiment analysis is now a relatively mature area of research, having received much attention from a number of research disciplines for more than a decade. Now we cover a few of the more important research challenges and some notable research works which address these challenges.

One early research problem which emerged in sentiment analysis was that of interpreting sentiment of user-generated reviews. As sentiment-bearing text, reviews have a

number of features which make them an ideal testbed for sentiment analysis. For one, reviews of products, films, music (etc.) have been freely available online since the days of USENET. Secondly, review text is, at least theoretically, holistically relevant to the topic in question. This means that document-topic relevance can be assumed, and no relevance determination is required. Thirdly, reviews are frequently accompanied with a sentiment annotation from the author ("4 out of 5 stars", "80%", "thumbs up") which may readily be used to provide a ground truth for evaluation. Lastly, and perhaps most importantly, the text is inherently subjective; the purpose of the text is to offer a subjective appraisal of the topic in question.

Two notable early works which deal with the problem of classifying reviews according to sentiment are Pang and Lee's work on movie reviews (Pang and Lee, 2004) and Dave *et al.*'s work on product reviews (Dave et al., 2003). Later works focused on the task of more granular sentiment analysis where the task was not only to identify sentiment towards topics, but also to identify towards which facets of the topic sentiment is expressed (Liu et al., 2005; Hu and Liu, 2004). This is sometimes referred to as *opinion feature mining*.

Other early work treated sentiment at a more fine-grained level by using lexical and syntactic features to model the sentiment contained in individual sentences or phrases. Tasks in this area have included identifying propositional opinion and the opinion holder (Bethard et al., 2004), determining the intensity of sentiment expressed in opinion clauses (Wilson et al., 2004) and understanding how syntactic structures and term prior polarity may be used to describe the sentiment of phrases (Wilson et al., 2005). These techniques have also been used to derive feature sets for document-level classification, for example using dependency trees (Matsumoto et al., 2005) or using appraisal groups (Whitelaw et al., 2005). Other related notable work includes that of Riloff *et al.* who use an information extraction approach to identifying subjectivity in text (Riloff and Wiebe, 2003; Riloff et al., 2003, 2005).

The aforementioned research for the most part does not however address ad hoc sentiment analysis scenarios, where the topic is not known in advance. This introduces two new challenges:

1. *topic heterogeneity*: topics may be very different in nature and thus techniques may suffer from domain transference problems.

2. *determination of relevance*: topically relevant textual content must be distinguished from non-relevant content.

Ad hoc sentiment analysis is a much more difficult task than for example review classification. For this reason, the problem is often formulated as an information retrieval task where approaches can rank documents for sentiment in a probablistic way, rather than as a binary classification. Approaches to this task include reformulating IR queries with opinionated words (He et al., 2008), classifying a document's individual sentences for opinionatedness (Zhang et al., 2007) and using document-level supervised learning (Gerani et al., 2009). These approaches each introduce sentiment components into a standard document retrieval model.

Another important task in sentiment analysis is that of aggregation. Given robust sentiment techniques, how may we summarise the sentiment at an aggregate level? One example which develops the aforementioned concept of feature mining, attempts to use extractive summaries to describe the sentiment towards a product's features (Hu and Liu, 2006). Ku et al. (2006) provide a more general application of information extraction methods for opinion extraction in news and blogs. Other research has tackled the summarisation problem by employing opinion source resolution (Stoyanov and Cardie, 2006).

A task related to summarisation is that of measurement. Although measurement also concerns the aggregation of sentiment over a body of content, the goal of measurement is to quantify the sentiment, often as a temporal series. Using quantitative methods allows us to measure sentiment in an opinion poll (O'Connor et al., 2010a), or to characterise debate performance (Diakopoulos and Shamma, 2010). We can also use such systems to track the overall mood online in blogs (Mishne and de Rijke, 2006a) or in news (Brew et al., 2010a). The latter is noteworthy for its bias correction in characterizing sentiment. A very exciting research topic at the moment is research into the predictive nature of these sentiment signals. Promising works have been completed for movie earning projections (Mishne and Glance, 2006; Asur and Huberman, 2010), for predicting the stock market (Bollen et al., 2011) and for political election outcomes (Tumasjan et al., 2010; Kim and Hovy, 2007). We have also recently completed some work on using Twitter to monitor political sentiment and predict elections results for the Irish General Election (Bermingham

and Smeaton, 2011) (in press).

So we see that sentiment analysis over the last decade or so has evolved into a field with many research topics and challenges. The reader is directed to the following texts for further reading:

- *Opinion Mining and Sentiment Analysis*: Pang and Lee provide a comprehensive review of research in the area of sentiment analysis up to 2008 (Pang and Lee, 2008).
- *Sentiment analysis and subjectivity*: The sentiment analysis portion of the *Natural Language Handbook* offers a thorough introduction to practical techniques and applications in sentiment analysis (Liu, 2010).
- *Computing Attitude and Affect in Text: Theory and Applications*: Shanahan *et al.* present a collection of works on the analysis of affect in text, covering a variety of approaches and applications in sentiment analysis (Shanahan et al., 2006).

2.2.4 Sentiment Evaluation Activities

Interest in this area of research may be attributed, at least in part, to the popularity of workshops dedicated to common research challenges. These challenges centralise the often considerable resources involved in performing large-scale evaluation, and provide common tasks for the research community. This provides an opportunity for researchers to benchmark and replicate experiments in a reliable manner. These workshops are often born out of a demand, or requirement, from commercial entities for solving a particular problem or progressing technology in a particular area. Many of the works discussed already in this chapter have stemmed out of these research activities. Here we give a brief overview of these tasks.

One such workshop is the Blog Track, which was introduced at TREC (Text REtrieval Conference) in 2006 (Ounis et al., 2006) and ran annually until 2010. The Blog Track focused on the challenge of *ad hoc search* of blog posts, finding relevant blog feeds (*feed distillation*), blog *opinion search* and *faceted blog distillation*. Other focuses included evaluating the potential benefit of *spam filtering* and identifying *top news* stories. The track issued participating groups with a common data set of blog posts, `Blogs06` (MacDonald

and Ounis, 2006). This was followed in later years by a much larger corpus, `Blogs08`. MacDonald et al. (2010) provide an overview of the Blog Track through the years.

As an example of the TREC methodology, let us look at the Blog Track opinion finding task which ran from 2006 to 2008. We participated in the Blog Track in 2008 (Bermingham et al., 2008) and were one of the top-performing systems for opinion finding and polarity detection (Ounis et al., 2008). Given a topic, participants were asked to find documents which were (i) *subjective*, (ii) *subjective and negative*, and (iii) *subjective and positive* towards the topic. A common participant approach was to rank 1000 posts for relevance for each topic and then re-rank these lists three times, each time ordering the documents according to opinionatedness, negativity and positivity. After participants submitted their runs, the results were pooled, and human assessors (or "annotators") labelled document-topic pairs with one of: *relevant*, *neutral*, *positive*, *negative*, *mixed* or *not judged*[2]. These labels are then used to calculate measures of retrieval effectiveness for individual result sets, such as mean average precision and recall. This allows systems to be easily compared in line with the Cranfield evaluation paradigm, which has been the dominant evaluation methodology in information retrieval for a number of decades (Cleverdon, 1967). This evaluation method conforms to the laboratory model for information retrieval evaluation (Saracevic, 2007b).

The data from the Blog Track was also used in the TREC TAC (Text Analysis Conference) 2008 Question Answering Track opinion question answering and summarization tasks (Dang, 2008). The Question Answering Track challenged participating groups to build systems to address two types of opinion questions: *rigid* questions and *squishy* questions. Rigid questions concerned more factual details such as "Who likes Mythbusters?" whereas squishy questions were more complex in nature e.g. "Why do people like Mythbusters?". Rigid questions are more amenable to precision and recall evaluation measures while squishy questions were evaluated using nugget pyramids, where multiple annotators are used to give higher weights to commonly interpreted answers.

Another similar activity, is the NTCIR (NII Test Collection for IR Systems) Multilingual Opinion Analysis Task (MOAT). Originally a pilot task in NTCIR-6 (Seki et al., 2007), this became a primary challenge in NTCIR-7 (Seki et al., 2008) and NTCIR-8 (Seki

[2]An annotator may wish to abstain from judging inappropriate content.

et al., 2010). The focus in MOAT was different from the TREC Blog Track in two primary ways: (i) a more fine-grained approach to opinion-finding was used, where the goals were to identify subjective clauses and opinion-holders, and (ii) tasks covered a number of languages: Japanese, English, Traditional Chinese and Simplified Chinese. A cross-lingual opinion question answering was introduced in NTCIR-8. They too evaluate using standard measures of retrieval performance and employ a methodology based on common data, annotations and tasks.

Of these, the most relevant task to us is the TREC Blog Track. One conclusion from the TREC Blog Track was that due to the inherently opinionated nature, strong ad-hoc retrieval systems with no sentiment-specific techniques performed well on the opinion-finding task (Ounis et al., 2008). This was found to be even more so the case the stronger the ad-hoc approach used, as opinion-specific system features produced slimmer margins of improvement over ad-hoc techniques for the opinion-finding task than systems with weaker ad hoc baselines. So, with a strong retrieval baseline, sentiment has arguably only a minor role to play in blog retrieval, even when the focus is a sentiment-based ranking. The implication is that it is the position of the non-relevant documents in the ranked list that is affecting the results rather than the sentiment of the relevant documents.

However, the notion of ad-hoc retrieval is very different in a microblog context, particularly for persistent search. Rather than a scarcity of relevant documents, often the issue is one of finding high-quality documents in an abundance of "relevant" documents. This new scenario with much fewer non-relevant documents offers a promising potential research avenue for sentiment analysis. Can analysing sentiment in these real-time microblog streams and identifying subjective commentary augment the persistent microblog search experience?

2.3 Evaluation Methodology

Having covered background to this work in terms of information retrieval and sentiment analysis, in the section we focus on the experimental methodology necessary to examine the role of sentiment in microblog persistent search. There are a number of important considerations in devising our experiments, including choice of experimental approach,

modelling human judgments and establishing methods which allow us to examine different aspects of sentiment. There are also experimental design considerations for evaluating the sentiment analysis portion of our system; this is covered separately in Chapter 4.

2.3.1 Static Corpus Evaluation vs. User Study Evaluation

An important decision must be made between an evaluation using a static corpus of documents, topics and judgments and conducting a laboratory user study evaluation. Evaluations using static data have several advantages including the high reproducibility of experiments and the comparability of systems. As we have already seen, this has been the dominant methodology in information retrieval in recent years, particularly in TREC and NTCIR workshops.

However, the new challenge of microblog search is fundamentally different, particularly for persistent search scenarios. No longer is the goal to identify messages which contain information relevant to the query topic; after all, in persistent search, topics frequently have many relevant topics which can be identified with high precision, for example by filtering using a hashtag, or a straightforward Boolean query. The more pressing task is to identify, to present to the information seeker, relevant documents on which they place a high value.

Secondly, static evaluations rely on the objective judgments of topic-document relevance (or some equivalent) by assessors to calculate metrics which describe a system's performance such as precision, recall and F-measure. Due to hindsight bias, these objective judgments are problematic to obtain. For example, if we now retrospectively look at the Irish General Election knowing the final outcome, we have a different perspective on the significance of content posted during the election campaign than we would have had at the time of posting. The same can be said for other real-time events: sports matches, television programmes, breaking news stories. Our assessment of the information utility at points in time after the initial information need is inherently influenced by a posteriori knowledge.

A third problem is that objective assessor judgments do not account for a user's internal knowledge, experience or outlook. We intuit that these factors are highly influential in

real-time information seeking. We therefore wish to incorporate these in our evaluation and not simply account for them using generalising assumptions.

As one recent review of search in microblogs concluded:

> *"...we should be strategic in crafting assessment methodologies at this early stage of research and development in microblog retrieval. Serious consideration of 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 echoes a sentiment proposed almost 20 years ago by Robertson and Hancock-Beaulieu of information retrieval:

> *"there has been increasing acceptance that stated requests are not the same as information needs, and that consequently relevance should be judged in relation to needs rather than stated requests. (A variant on this theme requires that relevance should be observed behaviourally, i.e. should be inferred from some action on the part of the requester.)"* (Robertson and Hancock-Beaulieu, 1992)

Collectively, these observations motivate our decision to evaluate our research hypothesis with a series of user studies. Our experiments are designed to capture user behaviour and use this as the measure with which we can evaluate the role of sentiment in our system.

2.3.2 User Study Design

Although static evaluations have been the focus of much information retrieval literature, user studies have been the primary method for evaluating information retrieval interfaces. These user studies typically focus on aspects of system usability which may help or hinder user performance in search tasks. Though interface evaluation is not an objective of our research, this type of evaluation provides an established foundation for our experiments. Our primary reference for our experimental design throughout this research is Chapter 2 of *Search User Interfaces* by Marti A. Hearst (Hearst, 2009). As a reference for the

statistical considerations of the experiments and evaluation we use *Statistics for Psychology* by Arthur and Elain Aron (Aron and Aron, 1999).

As Hearst notes, there are two main types of search usability studies: informal studies and formal studies. Informal studies are where participants are observed and interviewed regarding their interaction with the search system or mock ups of potential designs. This type of "user-centred design" is of particular use in the formative stage of design, when there are many possibilities and the design has not converged. In formal studies and controlled user experiments, users are exposed to variations of a system configuration to determine the effect that various factors have on system performance. We are certain about the design and variables we aim to examine, and thus, it is the latter style of experiment that we use in our studies.

In order to describe our experimental variables, let us consider a real-time, microblog, persistent search scenario. A user describes their information need to a system, say by providing a hashtag for a breaking news story as a query topic. The system then uses a relevance criteria to filter the stream and documents which satisfy this criteria are presented to the user in reverse chronological order. As and when new relevant documents become available, they are prepended to the list. This continues until such a time that the user determines that their information need has been fulfilled, the topic gains few new relevant documents or the user must abandon the search for some reason.

With respect to sentiment, we may characterise three aspects of this scenario with respect to a given topic:

1. *Document-level*: The sentiment contained in the content of each individual document towards the topic.

2. *Stream-level*: The distribution in document-level sentiment for a stream of documents.

3. *User-level*: The user's own sentiment towards the topic.

It is these three levels of sentiment we wish to investigate and constitute three of our independent variables. Other factors which are likely to influence task performance are those related to the users themselves. For this reason 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 can expose sets users to different stream-level sentiment (and hence document-level sentiment) at different times. In order to draw conclusions and comparison, we require a method for measuring the perceived quality of each configuration. This is achieved by capturing user system feedback. This user feedback is therefore our dependent variable and allows us to measure the effect which sentiment has in a microblog search and also evaluate with respect to a variety of user attributes.

It is important in these types of experiment to prevent biases such as order effects, learning effects and user contamination. As we describe in our evaluation chapters we are careful to use Latin squares and other randomised blocking techniques to mitigate these effects. At no time before or during experiments was the true nature of the evaluation disclosed to the participants.

2.3.3 Feedback

The notion of relevance is an important concept in information retrieval and information science. Particularly in information science, there has been much research into relevance's various complex manifestations and effects. Saracevic provides us with an in-depth review of information science research concerning relevance (Saracevic, 2007a,b). However, as we have already explored, a corpus of objectively relevant document is not appropriate for our evaluation; our evaluation relies on user feedback.

In order to determine how we might implement feedback in our experimental system, let us consider the requirements for such a mechanism:

- *Real-time*: Documents must be judged shortly after they are written.
- *Non-intrusive*: System feedback must be made by participants with minimal effort so it does not usurp unnecessary time and detract from the user's primary task of search.
- *Intuitive*: The nature of the feedback must be consistently and easily understood by all experiment participants.

- *Discriminative*: The feedback mechanism used must enable users to disciminate clearly between documents they perceive as valuable given their query, and those they do not.

These criteria can be met with an inline real-time system feedback function. Users can provide feedback for documents as they appear in their stream and which is then stored for later analysis. At certain stages, we will require feedback from the users which assesses the overall quality of a stream of documents. We therefore must prompt the users for feedback immediately after they have experienced a given stream configuration. The user action required to give the feedback must be non-taxing so that the task is non-intrusive; we must only capture feedback absolutely necessary for our evaluation and in a manner that minimizes cognitive load on part of the user.

The other two constraints relate to the definition of the feedback itself. Users must be accurate and comfortable in their feedback. For this reason we use familiar UI patterns accompanied by clear and concise instructions and training. The feedback must also be granular enough to allow users to express themselves efficiently and accurately, yet simple enough that we may use it to perform useful statistical analysis.

2.4 Conclusion

In this chapter, we have given a detailed overview of the research problems and methodology associated with this research. We have also examined the related research from the fields of information retrieval and sentiment analysis and in doing so provided motivation for our hypothesis. We discussed how persistent microblog search can be modelled as an information filtering task. Finally we examined the criteria for evaluating our research system and derived a methodology from information retrieval interface evaluation literature.

It is clear that evaluation methodology for information systems needs to evolve if we are able to reliably measure system performance in a real-time context. It is intended that the methodology we develop here is a step towards establishing a new temporally-focused evaluation methodology where real-time feedback is the judgment necessary to experimentally measure the relative successes and failures of various system configurations.

Most pertinently, this methodology will enable us to accurately measure the effect that sentiment has in real-time microblog search.

Having established our methodology at a high level, we next seek to operationalise this methodology in a real world system. Before conducting our experiments, we must specify a system with satisfies the requirements set out in this chapter for our experiments. In the next chapter we detail our experimental system architecture, describe our user scenarios and define our evaluation measures.

Chapter 3

A System for Examining Sentiment in Real-time Microblog Search: *Channel S*

Before proceeding with our evaluation, it is necessary to consider the design and development of the system which will support our experiments. After all, as with any technological evaluation, the capabilities and limitations of the underlying technology define boundaries for the experimental design. For our experiments, we have designed and implemented a system called *Channel S*, a real-time system for *S*earching with *S*ocial *S*entiment.

In this chapter we consider the development of the system from a number of perspectives. In Section 3.1 we describe the architecture of the system. In Section 3.2 we describe the user interaction and how this is incorporated into the system interface. In Section 3.3 we discuss how the system supports our methods and measures for experimental evaluation and we conclude the chapter in Section 3.4.

3.1 System Architecture

Channel S is architected as a real-time web-based system. The architecture is componentised and loosely coupled so that we may conduct evaluations at a subsystem level. In this section we detail the system requirements, data and implementation.

3.1.1 Functional Requirements

Let us consider at a high level the primary requirements of the system:

1. *Real-time*: The system must facilitate real-world, real-time microblog search tasks.
2. *Controlled Sentiment*: The system must allow microblog sentiment to be controlled.
3. *Natural*: The system must look and feel similar to other microblog search systems to mitigate learning effects.
4. *Feedback and Evaluation*: The system must record user interactions which support our evaluation.

These requirements each deserve attention at design stage. In the case of (1), we ideally wish to evaluate with as many diverse real-time topics as possible. However, given limited participant and laboratory resources it proves unfeasible to run a large amount of real-time laboratory user trials. Our design allows us to run a user study with simulated real-time data to give broad topic coverage, before running real-time user studies to explore certain topics in much greater depth.

For (2), controlling sentiment requires that our system contains a module which can accurately and efficiently determine the sentiment of content towards a given topic. A sentiment analysis module in a system must be treated with caution; sentiment analysis systems are far from perfect and make many incorrect sentiment decisions. Microblogs as a new area of study must be especially approached with caution. Sentiment analysis techniques which have a proven track record on much longer and more well-structured text may not transition well to microblog content. Requirements (3) and (4) are discussed later in this chapter when we consider our feedback interaction and our experimental measures.

For these reasons, in this thesis work we do not proceed to deploy a fully-automated real-time sentiment-based search system initially. In our first experiment, we evaluate the sentiment analysis module outside the context of a search scenario (Chapter 4). In our second evaluation, we simulate real-time search tasks using human-judged sentiment (Chapter 5). This allows us to examine multiple topics and exert precise control over sentiment without relying on automated sentiment classifications. In doing this we can

understand the performance of our system and make initial observations before we conduct our live real-time experiments (Chapter 6).

3.1.2 Data

There are two types of data we require for our experiments: content and topics. We have chosen to use the microblogging service, Twitter[1], throughout our experiments as our content source. Twitter has come to define microblogging, particularly in terms of the follower model of social connection and 140 character maximum post length. Other notable microblogging platforms include the widespread social network, Facebook[2], enterprise microblogging platform, Yammer[3], and the Google-owned, open source Jaiku[4], and more recently, Google+[5]. However, none of these can currently compete with Twitter in terms of availability of data and associated programming interfaces, as well as user base and mainstream prevalence.

Our search topics throughout are modelled on Twitter trending topics. In doing so, we make a reasonable assumption that these trending topics approximate topics of interest on Twitter. That is not to say that there are no other topics of interest on Twitter which are dissimilar to trending topics. However, identifying and accomodating such topics is outside the scope of this work.

3.1.3 Implementation

Channel S may be broken down into a number of components:

- *Relevance Filter*: Connects to the Twitter streaming API (Application Programming Interface) using a third party library[6], and filters for relevant content.
- *Topic Descriptors*: Provides data to the relevance filter and the sentiment analyser about the topic and associated sentiment targets.

[1] http://www.twitter.com
[2] http://www.facebook.com
[3] http://www.yammer.com
[4] http://www.jaiku.com
[5] http://plus.google.com
[6] http://twitter4j.org

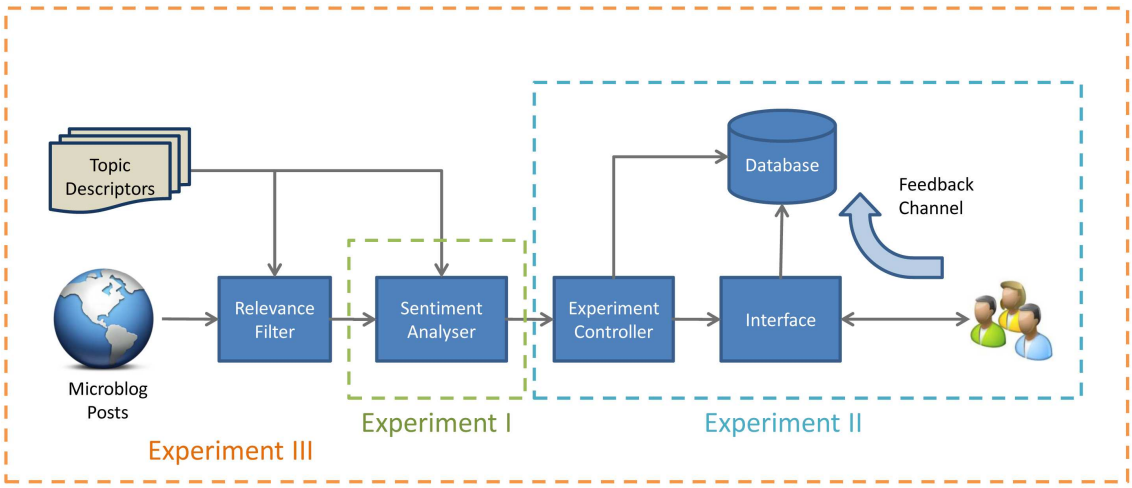


Figure 3.1: Conceptual diagram of the *Channel S* system.

- *Sentiment Analyser*: Analyses sentiment in content with respect to sentiment targets. Contains text pre-processing pipeline, feature extraction and trained classifier.
- *Experiment Controller*: Controls experimental variables, provides content to presentation layer and prompts users for survey feedback when required.
- *Interface*: Web presentation and interaction layer for displaying microblog post stream and notifying users when they are to give feedback.
- *Feedback Channel*: A service which allows feedback to be uploaded to the database in the background via interface actions.
- *Database*: Stores all content, sentiment classifications and user feedback for later analysis.

The system is illustrated conceptually in Figure 3.1. This diagram also indicates the portions of the system which feature in each experiment. As mentioned previously, for our simulated experiment, the relevance filter and sentiment analyser are replaced with a corpus of human-labelled documents. In our machine learning evaluation, the sentiment analyser is considered in isolation.

3.2 Interaction Design

Regarding user experience, our goal is to make the interaction as unintrusive and as familiar as possible. User studies for real-time search are a new challenge and, as such, we strive to minimise the risk of causing adverse affects by introducing unfamiliar elements. In this section we discuss the considerations in designing our system interface. Before deploying the system, two pilot testers were observed using the system and asked to give feedback. Their comments while using the system, and in informal interviews afterwards, were used to refine the system interface.

3.2.1 Interface

The interface is designed as passive so the stream is constantly receiving and displaying new posts without user action. This means that, for example, if a user's focus is diverted from the stream, they can look back and catch up at their leisure. This pattern is that observed for searches in the popular, Twitter-owned desktop client, TweetDeck[7]. See Figure 3.2 for comparison screenshots of persistent Twitter search in TweetDeck and Channel S.

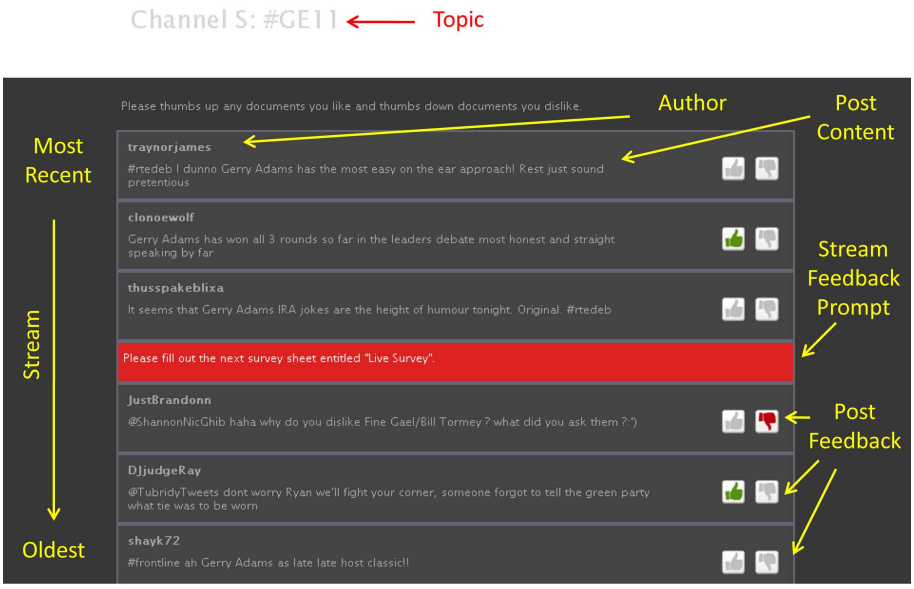
We use a combination of technologies to provide this web interface. We use PHP[8] on the server side and the powerful JavaScipt library JQuery[9] to present the stream to the user on the client side. The user's actions (feedback) are sent to the server via JQuery AJAX commands and then stored in a database for later analysis.

The primary visual element is a series of microblog posts in descending chronological order. New microblogs posts are periodically prepended to the list and older posts shift down accordingly. Initially we intended to simply display the content of the posts, but in pilot tests we found that users were not comfortable not knowing the name of the author when judging the content. We speculate that this is due to a level of context offered by the information in the Twitter username; the username for example could be used to determine an author's gender, differentiate between a personal and a company account or to help identify spam.

Our sentiment analyser processes posts in batches. For this reason, the experiment

[7]http://www.tweetdeck.com
[8]http://www.php.net
[9]http://jquery.com

(a) Persistent Twitter search in TweetDeck.



(b) Channel S uses a familiar persistent search presentation, with feedback elements integrated.

Figure 3.2: A comparison of search streams on Channel S and Twitter client, TweetDeck.

controller retrieves new content also in batches. The controller drip feeds posts to users at a rate of one every 10 seconds (an interval we tuned with users during pilot testing). There is therefore a latency in our system from authoring to presentation equal to the sum of the crawling time, sentiment processing time and presentation time. During development, we ensured that there was never a latency of longer than 60 seconds, a period we deemed acceptable for our purposes. It should be noted that there is no reason that this could not be reduced to a few seconds, or indeed subsecond latency, with sufficient resources.

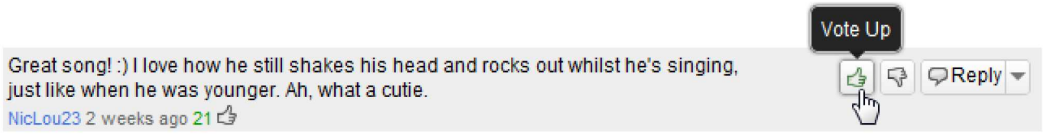
3.2.2 Feedback

There are two types of user feedback we require from our system. First we need to record a user's assessment of individual posts and secondly, we must record user assessment after viewing a stream of documents for a period of time. We now discuss each of these in turn.

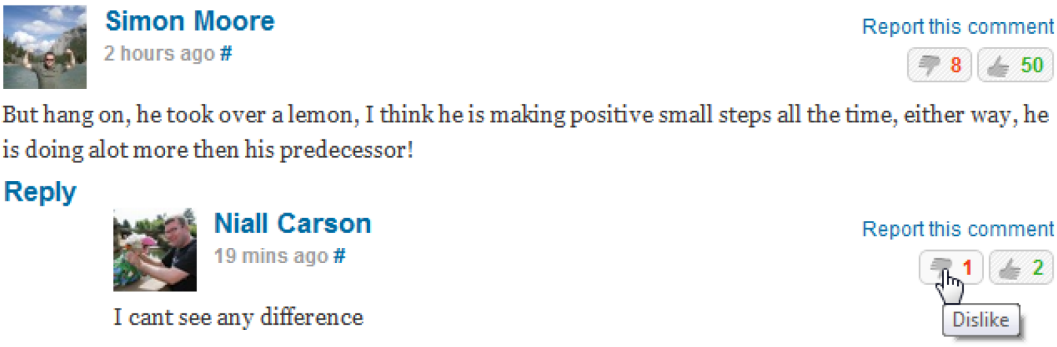
As discussed in Chapter 2, rather than using an objective concept like relevance, we wish to record a measure which more behaviourally reflects a user's assesment of the value they perceive in microblog content. We must define a feedback mechanism which complies with the requirements we have outlined. To do this, we adapt an established interaction metaphor found throughout social media and the modern Internet, *thumbs up* and *thumbs down*, sometimes referred to as *like* and *dislike*. See Figure 3.3 for examples. Allowing users to annotate microblog posts in this way means that for each document presented to a given user, it will have received one of three annotations: *thumbs up*, *thumbs down* or *no annotation*.

But how should we define these feedback actions for users? We instructed users to approach liking and disliking a document as they would if they enountered it in their normal Internet use. They were told that the annotations they give were to be used to improve performance in a new real-time microblog system. They were not told of the true sentiment-oriented focus of the evaluation. The guidelines in full can be read in Appendix C.

As well as extrapolating a measure for overall stream feedback from individual document annotations, we also wish to prompt the user to provide an explicit evaluation of a series of documents at a stream level. This allows us to capture a user's perceived utility



(a) Video comment on YouTube (http://www.youtube.com)



(b) News article comment on thejournal.ie (http://www.thejournal.ie)

(c) Answer on community question-answering service, Quora (http://www.quora.com)

Figure 3.3: Examples of content feedback in social media.

for a period of time during the search task where their stream was assigned a given filtering algorithm. Where required, users are asked on a 5-point Likert agreeability scale whether they thought the preceding stream was each of:

- *Interesting*: User assessment of whether the content in some way evoked their interest, or was intriguing.
- *Insightful*: User assessment of whether they found the content offered some unique insight or point of view.
- *Informative*: User assessment of whether the content in the stream was providing them with new, relevant information.

Users are also prompted to provide an overall rating for the preceding stream on a scale of 1 (poor) to 7 (excellent). We regard this as our primary stream-level feedback and the previous three dimensions as secondary feedback. Together, the primary and secondary feedback measures enable us to reason about user motivations.

3.2.3 User Profiling

Naturally, different users will approach this task differently and, as we discussed in Chapter 2, user characteristics can have a significant impact on user behaviour in a search task. There are three aspects of user profile we wish to capture: *task familiarity*, *demographics* and *a priori topic sentiment*. For task familiarity, we ask the user about their familiarity with Twitter and its various features such as posting, reading and most importantly, searching. For demographics we record gender, age and level of education. Although not an extensive demographic profiling, this allows us to make some observations about what effect, if any, demographics has on observed user feedback. For a priori topic sentiment we record the user's stated sentiment towards the topic, and where applicable, entities related to the topic. We anticipate that a user's internal set of beliefs might have a significant bearing on how they perceive sentiment which may be aligned with or against their own personal sentiment. We also allowed users to declare themselves as either being unfamiliar with the topic or being familiar, yet having no strong sentiment, to cater for all scenarios.

3.3 Experimental System Configurations

Our research consists of a series of three evaluations: (i) an evaluation of the supervised sentiment classifier, (ii) a simulated real-time user evaluation and (iii) a series of live real-time user studies. Each of these requires a different system configuration and different evaluation measures. In this section we describe how Channel S supports each of these experiments in turn.

3.3.1 Experiment I: Sentiment Analysis for Microblog Posts

In this experiment, we isolate our supervised learning sentiment analysis component. It is well understood how to evaluate such systems. We first develop a body of labelled data with known sentiment. After identifying candidate feature sets and classifiers, we perform a series of 10 fold cross validation tests on the labelled data. By using cross validation, we use the body of labelled documents to simulate unseen test data. Our primary metric for performance is classification accuracy and we can use this to benchmark against sentiment

classification for other textual domains and applications.

3.3.2 Experiment II: Simulated Real-time Microblog Search User Study

In our second experiment, our goal is to simulate a range of topics with short, simulated, real-time search tasks. This requires the full system interaction layer with an experiment controller which presents a user with a series of diverse topics and streams of relevant posts. Rather than a real-time web connection however, we use human-labelled data. This allows us to control the sentiment in the streams with a high degree of precision. For our evaluation we must capture user profile information, document-level feedback, stream-level feedback and user-topic sentiment.

3.3.3 Experiment III: Real-time Microblog Search User Studies

In our final evaluation we amalgamate the systems from Experiments I and II and use live real-time data. Unlike the configuration for the simulated task, this configuration is real-time and multi-user and as such, must be tested for load and latency. As well as real-time data, the system also requires prior labelled data to train the sentiment analyser. It would be possible to use the data from previous experiments but we chose to develop topic-focused training data as this is likely to yield a better performance.

The data capture necessary is similar to that for the previous experiment except for a few differences. In this experiment we have fewer topics but we capture the a priori sentiment in more detail. Secondly, the user must give stream-level feedback in real-time, so the system prompts users to complete the questions in hard-copy without interrupting the search stream. In this task, the volume of data viewed by the user is significantly higher.

3.4 Conclusion

In this chapter we have given a technical overview of our system for studying sentiment in real-time microblog search, Channel S. We have described its design and implementation with a specific focus on user interaction and system feedback. We have detailed the architecture of the system and described the function of the constituent components and required inputs and outputs. We have also described how this system is configured for each of our experiments and how the system captures the data necessary for our evaluation.

It should be clear at this stage that our experiments each in turn allow us to move incrementally towards evaluating our hypotheses. It is tempting to build our system and immediately deploy real-time user studies. As a new area of study however, with many poorly understood aspects, we feel it is vital to understand the constituent system components and technology before performing our final evaluation. In the following three chapters we present each of our evaluations in turn, culminating in real-time user studies conducted during live broadcast television events.

Chapter 4

Sentiment Analysis for Microblog Posts

Automated sentiment analysis is at the core of the research developed in this thesis. This sentiment analysis must be accurate and efficient if we are to employ it effectively in real-time during our user studies. In this chapter, we evaluate the appropriateness of machine learning methods for identifying sentiment in our chosen data: microblog posts.

Microblogs, as a new textual domain, offer a unique proposition for sentiment analysis. Their short document length suggests any sentiment they contain is compact and explicit. However, this short length coupled with their noisy nature can pose difficulties for standard machine learning document representations. In the following sections we examine the hypothesis that it is easier to classify the sentiment in these, short-form documents, than in longer-form documents. To do this, we developed a corpus of sentiment topics and document annotations from the popular microblogging service, Twitter. Using these annotations, we train classifiers and evaluate a number of document feature representations for sentiment classification. We also perform the same set of experiments on a collection of microreviews, and draw comparison and contrast between performance on these two, short-form domains, with two long-form domains: a collection of movie reviews and a collection of blogs. We achieve a higher accuracy in classifying sentiment in microblogs than in blogs. However, we find the opposite to be true for reviews and their short-form counterparts, microreviews. We observe also that ad-hoc sentiment classification is in

general a significantly more difficult task than review classification. Throughout we make a number of observations specifically pertaining to the challenge of supervised learning for sentiment analysis in microblogs.

In Section 4.1, we place our research in context of related work in this area. We follow this in Section 4.2 with a description of our methodology. The results of our evaluation and discussion are presented in Section 4.3, and we conclude the chapter in Section 4.4.

4.1 Background and Related Work

The short length of microblog posts means they can easily be published and read on a variety of platforms and modalities. This brevity constraint has led to the use of non-standard textual artefacts such as emoticons and informal language. These are often referred to as *sociolinguistic* features. The resulting text is often considered "noisy". Table 4.1 contains examples of microblog posts from Twitter. Note how the content ranges from well-formed sentences to more speech-like disconnected utterances or phrases, with frequent disregard for punctuation or grammar. Prevalent also are the use of emoticons (":-)", ";-D"), abbreviations ("b/c") and unconventional syntax ("*joy*", "right!?!"). There are also platform-specific features, such as hashtags which are used to denote a relevant topic ("#6Nations"), and usernames ("@afranks"). We also see more general Internet conventions such as square brackets to indicate source or content type ("[TechCrunch]") and URLs ("`http://ow.ly/4b9j`"). Clearly microblog content is very different in nature to conventional, well-formed, grammatical English text.

It is reasonable to assume that the short document length introduces a succinctness to the content. The focused nature of the text and higher density of sentiment-bearing terms may benefit automated sentiment analysis techniques. On the other hand, it may also be that the shorter length and language conventions used mean there is not enough context for sentiment to be accurately detected due to the sparse feature vectors. It is unclear which of these is true.

These issues motivate our research questions — recall from Chapter 1:

1. *In what ways do the natural language and the textual conventions used in microblog text differ from that used in other types of user-generated content?*

Dear friends...I have a problem(admitting it is the first step right!?!)A sample of my Twilight paraphernalia http://twitpic.com/2di4u *joy*
correction: Scotland Women 10 - 31 Wales Women. #6n #6Nations
Safari 4.....wwwwwwoooooooowwwwwww
NO GOLF TODAY TOOOOO COLD....guinness crackin pizza delivery time! BOOYAH
Rolling home after a long days work. Currently reading Child 44 on my kindle for iPhone.
didn't get the chance to say hi to @mitchfree but the guy looks like he used to rock alongside Bono or something. ;-D
In 1992, the oracle of Omaha predicted the decline of newspapers, magazines, and TV. And not b/c of the internet. http://ow.ly/4b9j
[TechCrunch] IBM Uses Amazon To Leapfrog Microsoft On The Way To The Blue Cloud http://tinyurl.com/dywgg6
@afranks Haha sure :-) At least I still get to go to coachella :-)
I'm probably the last to find out about this "Wolfram Alpha" really intelligent software threatens google

Table 4.1: Examples of microblog posts from Twitter

2. *What effect does the nature of microblogs have on sentiment analysis using supervised learning for microblog posts compared to traditional, longer document classification? What comprises an optimum feature set and classification strategy?*

4.1.1 Microblogs as a Noisy CMC Domain

Microblogging, like many other Computer-Mediated Communication (CMC) platforms, exhibits a higher level of noise compared to print domains. A number of studies have described the nature of new CMC domains. Mishne discussed the nature of blogs in relation to the British National Corpus, a standard corpus of English language documents (Mishne, 2007). Mishne also examined a range of text analytical approaches for blogs, including information retrieval and sentiment analysis. Herring *et al.* attempted to describe blogs as a genre, exploring antecedants and the language use exhibited (Herring et al., 2004). They conclude that blogs are a hybrid of other genres, and also note that although the technology trigger for the advent of weblogs was relatively small, they tend to have a comparitively high impact as a communication medium.

A perception of social content is that informal language and textual artefacts are commonplace. Tagliamonte and Denis studied the language used by teenagers in instant messaging (Tagliamonte and Denis, 2008), finding that instant messaging was a hybrid genre and that many of the traits we associate with noisy CMC text ("lol", ";-)", "OMG!!!!", etc.) tend to be used less in adolescents as they approach the end of their teens. They also concluded that the penetration of non-standard English language and punctuation is far less than is reported in the media. In a study of classification of customer feedback, Gamon found a high level of accuracy for supervised sentiment classification despite their noisy nature (Gamon, 2004). Hård af Segerstad described in detail the linguistic nature particular to a number of CMC domains: email, web chat, instant messaging and SMS (af Segerstad, 2003), finding commonalities and uniques features in each domain. Carvalho *et al.* found that non-standard surface features such as a heavy punctuation and emoticons are key to detecting irony in user-generated content (Carvalho et al., 2009). One study has looked specifically at word-lengthening (e.g. "cooooolllll!!!!") and has proposed a method for modelling such spelling variants for sentiment detection (Agarwal et al., 2011). In-

terestingly, they find that approximately one in six microblog documents contains word lengthening of some kind, and that the more likely a word is to be lengthened, the more likely it is to be a sentiment-bearing term.

One strategy to accomodate non-standard language put forward by Choudhury *et al.* is to use Hidden Markov Models to decode text into standard English (Choudhury et al., 2007). Choudhury's work reports a high rate of success in normalising SMS (text messages). More recently, an unsupervised method for normalising ill-formed words in microblog content has achieved promising results both on microblog text, and on SMS text (Han and Baldwin, 2011). In this work, the authors report that over 15% of microblog documents contain more than 50% out of vocabulary terms. Agarwal *et al.* showed that by simulating noise in text classification, a good classifier should perform well up to about 40% noise (Agarwal et al., 2007). This suggest that, although noise may be present in text, this may not prove to be important for supervised learning tasks. Foster et al. (2008) investigated adapting parsers to noisy text data, finding that they are able to adapt parsers trained on print corpora to grammatically noisy corpora without affecting performance on grammatically well-formed text.

Using text features derived from parsing text has shown promising results for sentiment classification, in particular Matsumoto's work on movie review classification (Matsumoto et al., 2005). More recently Foster *et al.* have turned their attention to parsing and POS tagging microblog content with promising results. It is worth mentioning some of the other recent works that have used syntactic features in sentiment analysis tasks. Wiegand and Klakow (2010) experimented with various kernels for extracting opinion holders from text, finding better performance from tree kernals than from vector or sequence kernels. They find that the best performance, however, is when all kernels are combined. Johansson and Moschitti (2010) demonstrated that using features based on syntactic and semantic structures can achieve a modest increase in performance for identifying subjective text on the MPQA corpus (Wilson et al., 2005). (Nakagawa et al., 2010) employed a similar approach and found syntactic structures better than bag-of-ngram models for sentiment tasks, both in English and in Japanese. (Karlgren et al., 2010) use *constructional patterns* such as "tense shift" and "verb chain" to model text in a way that is not relient on lexical information, but which is not as involved as using tree-based features. They argue that

their approach is effective, suitable for mitigating the effects of domain transference and is relatively low cost. (Wu et al., 2009) use dependency parsing to identify opinion holders, product features and the opinion expressions linking the two. They find that an SVM using a tree kernel outperforms SVMs which use combinations of lexical, POS, ordering, distance or binary dependency relation features. Some works are starting to appear in the microblog domain (such as Agarwal et al. (2011)) which make use of syntactic trees and POS tags for sentiment feature engineering.

As noted recently, there is still a significant challenge in adapting existing parsing and POS tagging techniques to microblog content, and Web 2.0 content in general (Foster et al., 2011). Thus, although these techniques have gain considerable traction, we feel it premature to rely on them for our evaluation. The computational considerations could also prove problematice for a system with a real-time contraint.

4.1.2 Sentiment Analysis for Microblogs

Some exploratory works have been reported on sentiment in the microblog domain. Diakapolous and Shamma used manual annotations to characterise the sentiment reactions to various issues in a debate between John McCain and Barack Obama in the lead up to the US Presidential election in 2008, finding that sentiment is useful as a measure to identify controversial moments in the debate (Diakopoulos and Shamma, 2010). Previously, Shamma *et al.* examined a variety of aspects of debate modelling using Twitter, beyond individual politician performance (Shamma et al., 2009). In these studies, Twitter proved to be an effective source of data for identifying important topics and associated public reaction.

Jansen *et al.* studied the word of mouth effect on Twitter wherein one of their focuses was how and why positive and negative sentiment towards brands was spreading (Jansen et al., 2009). Sentiment was classified using Summize[1], an adjective-based sentiment classifier for Twitter. They found their approach useful for analytics for brands in Twitter. Bollen et al. have focused on modeling public mood on a variety of axes to correlate with socio-economic factors (Bollen et al., 2009). They report a number of interesting observations such as changes in tension and anxiety around important events and find a

[1]no longer publicly available

significant improvement in predicting the Dow Jones Industrial Average when incorporating sentiment. This work is echoed by preliminary work from Zhang et al. who also focus on emotive concepts, in this case "hope" and "fear", and correlate with a number of market indicators (Zhang et al., 2010).

Techniques have been used in text classification to mitigate the effect of feature sparseness in short documents for classification. Metzler *et al.* used query expansion techniques and language modelling to expand short sequences of text, in their case queries, into longer segments of text in order to assess text similarity (Metzler et al., 2007). Healy *et al.* used a combination of word and document statistic features to help classify short texts for spam (Healy et al., 2005). Gabrilovich and Markovitch used real-world knowledge via ontologies to expand text (Gabrilovich and Markovitch, 2005). Although these all represent interesting work, there is very little work on the specific challenge of classifying short-form documents from the social web, like microblog posts; the majority of literature is devoted to much longer text classification.

These studies confirm our assumptions that microblogging and other similar domains are intrinsically different in nature from traditional text domains. The prevalence of noisy text and the degree to which it affects text categorisation however remains an open question. To our knowledge, this is the first work to explore the challenges that the shortness of microblog documents present to feature vector representations and supervised sentiment classification.

4.2 Methodology

In this section, we detail our methodology for constructing our dataset and corpus of annotations. We follow this with a standard machine learning evaluation for binary and three-way sentiment classification.

4.2.1 Developing a Microblog Corpus

The microblog posts used in these experiments are taken from a collection of over 60 million posts which we gathered from the Twitter public data API[2] from February to

[2] `http://apiwiki.twitter.com`

May 2009. Were also gathered *trending topics* on Twitter during this time. At any given time, trending topics are the most discussed topics on Twitter. We examined the trending topics and identified five recurring themes: *Entertainment*, *Products and Services*, *Sport*, *Current Affairs* and *Companies*. Assuming that these are representative of topic categories of interest on Twitter, we selected 10 trends from each of these categories to be used as sentiment targets, giving 50 topics in total. The posts we used for annotations were identified by looking in our collection for posts which mention each of the topic terms in any order. The full list of topics can be found in Appendix D.

Creating a diverse set of topics to be used as sentiment targets makes classification more difficult, as vocabulary and style vary from one topical genre to the other. Indeed, domain-specific classifiers would likely provide more accurate classification than the generic classifier presented here. Our topics include such diverse topics as "*The Afghanistan War*" and "*Susan Boyle*". By making the topic set diverse and challenging, we hope to better test the performance of our approach, and build a classifier representative of a real-world, generic, sentiment classification scenario.

In the annotation process, we used Wilson's definition of sentiment:

> *"Sentiment analysis is the task of identifying positive and negative opinions, emotions, and evaluations."* (Wilson et al., 2005)

Our team of annotators consisted of 9 PhD students and postdoctoral researchers with varying degrees of familiarity with sentiment analysis. To encourage agreement among the annotators, the annotation process was preceded by a number of training iterations. For the first round, we provided a draft set of annotation guidelines to the annotators. These guidelines outlined the annotation process, the annotation classes, topic definitions and gave examples of the three types of sentiment according to Wilson's definition: *Opinion*, *Emotion* and *Evaluation*. For each document that an annotator annotates, they must assign a sentiment category to the document, reflecting the document sentiment towards the assigned topic. As in Wilson's experiments, the annotators were asked to interperet the sentiment of the document as a whole, rather than deconstructing the text. The annotation categories were derived from our previous annotation work (Bermingham and Smeaton, 2009; O'Hare et al., 2009). Details of these categories can be found in Table 4.2.

After annotators had studied the guidelines, we asked them to label ten Twitter posts for each of five topics. Each annotation (topic-document combination) was performed by at least four annotators.

After the annotations had been collated, all annotators met as a group to discuss contentious annotations. A list of consensus annotations was drafted, the guidelines were updated to clarify ambiguities which were raised and the process was repeated with different topics. This yielded only a marginal increase in inter-annotator agreement. We speculated that this was due to a disproportionate influence on the consensus annotations from a few annotators during group discussions. Before the third round of sample annotations, each annotator participated in an individual training session where they were asked to annotate aloud and referring to the guidelines for their reasons for annotating. Again, we addressed ambiguities which were raised in revised guidelines. In the individual training sessions, it was apparent that annotators were considerably less clear on the guidelines than they reported in the group scenario. In our third and final iteration, we observed a significant increase in agreement among annotators. The annotator guidelines may be found in Appendix B.

The annotators chiefly reported three issues. The first was the *definition of sentiment*. The annotators reported that opinion-style sentiment was often easy to detect, however, evaluative or speculative sentiment proved more troublesome. Sometimes, factual statements can appear to be evaluating a subject in a positive or negative light. For example, it is difficult to interpret what is meant when an author reports a team winning a football match or reports on negative press towards a company. In these cases, is the author expressing an evaluation, or simply reporting fact? If annotators were unsure they were encouraged to use the *unclear* label to indicate that they are having trouble identifying the sentiment or that they feel they do not have enough information to make a sound judgement. As a rule of thumb, annotators were advised that if they spent longer than 30 seconds deliberating over a single annotation, they were unlikely to be able to annotate with a degree of confidence, and should indicate they are *unclear*. The annotation categories were derived from our previous annotation work (Bermingham and Smeaton, 2009; O'Hare et al., 2009). Full details of the annotation classes can be found in Table 4.2.

The second issue from our annotations concerned *topic definition*; annotators reported

it difficult to define the boundaries of some topics. If a post mentioned a topic only tangentially, should this document still be considered as a whole relevant to the topic? This is a problem which has plagued information retrieval for years leading to the development of graded relevance measures such as Discounted Cumulative Gain (Järvelin and Kekäläinen, 2000). For these experiments, we made a simplifying assumption; a document is considered relevent if the topic as it is defined in the topic description is referenced in the document.

Lastly, the issue of *topic-sentiment boundary* was frequently raised. For example, if a document expresses sentiment towards a player on a team, or a representative of a company, does this sentiment reflect on the team, or company, as a sentiment target? This topical ambiguity is frequently a problem for information retrieval relevance definitions. For a thorough treatment of the issues surrounding relevance in information retrieval, see Saracevic (2007b). For the purposes of these experiments, we adopted a sum-of-its-parts approach to topic definition. If a topic is a team, sentiment towards a player may be interpreted as sentiment towards that team. On the other hand, if the player is the sentiment target, sentiment towards a team does not indicate relevant sentiment.

We provided the topics in four parts: a topic title, a relevance guideline, a sentiment guideline, and a topic description. The topic description consisted of the first paragraph of the topic's Wikipedia article. This was provided to give the annotators context if they were unfamiliar with the topic.

We developed an annotation tool to be used by the annotators (see Figure 4.1). The annotation tool presents the document-topic pairs to users in batches of 50, allowing the user to pause or adjourn the annotation session between batches. The tool interface displays the topic details alongside the document and labels. We instructed the annotators to keep a printed copy of the annotation guidelines available at all times for reference.

In total, 9 annotators annotated 17 documents for each of the 50 topics giving 850 documents per annotator, 7,650 annotations in total. Annotations from the training iterations were discarded. One document per topic was also annotated by another annotator. In total, 463 documents (6.78%) were doubly annotated for testing inter-annotator agreement. For agreement across the 7 classes we observed a Krippendorff's alpha of 0.56 (or Fleiss's kappa of 0.65). If we consider just the 3 classes which will be used for training, *positive*, *negative* and *neutral*, conflating the remaining classes to *other*, alpha rises slightly

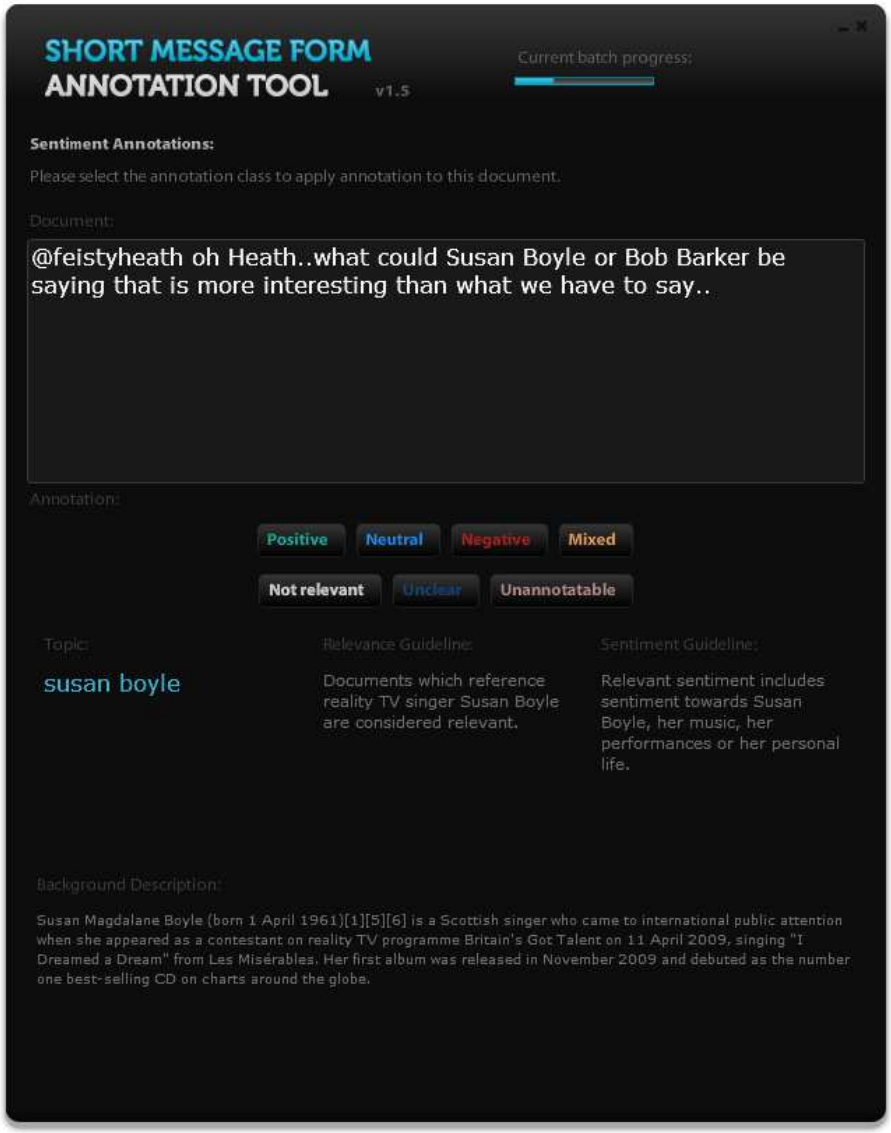


Figure 4.1: Sentiment annotation tool interface

Label	Definition	#Documents
Relevant, Positive	Predominantly positive towards topic	1,410
Relevant, Negative	Predominantly negative towards topic	1,040
Relevant, Neutral	Relevant to topic but no sentiment towards topic	2,597
Relevant, Mixed	Positive and negative sentiment towards topic	146
Not relevant	Not relevant to the topic	498
Unannotatable	Spam, Inappropriate, Non-English, etc.	603
Unclear	Not enough information to annotate or I am unsure	530
Total		6,824

Table 4.2: Microblog annotation labels and associated document counts

	Positive	Negative	Neutral	Mixed	Unclear	Unanno	Not Rel
Positive	62	3	25	6	11	3	5
Negative		38	18	3	13	2	9
Neutral			124	2	20	5	16
Mixed				5	2	0	0
Unclear					11	4	6
Unanno						35	7
Not Rel							28

Table 4.3: Matrix of pairwise inter-annotator agreement per label

to 0.58. If we just consider the binary sentiment classes, *positive*, *negative* and *other*, we get an alpha of 0.57. These results are consistent with our previous work in blog annotations (O'Hare et al., 2009). The encouraging values for alpha we observe mean we can rely on our training data with a degree of confidence. Although this is marginally lower than the suggested threshold for acceptable agreement presented by Krippendorff (0.67), Krippendorff also suggests that different tasks require alpha to be interpretted appropriately. In our challenging task of identifying sentiment, we consider our observed alpha to be sufficient for use in our experiments.

As one recent sentiment study noted of preparing labelled sentiment data for classification tasks:

> *"...if there is good agreement between annotators, then annotation effort should be expended on maximizing coverage rather than identifying consensus."* (Brew et al., 2010b)

The agreement we observe is sufficient that we can be confident in our document labels and there is no need, for example, to assign multiple annotators per document and only use consensus annotations.

An interesting point to note is that the above study also makes use of two promising methods for maximising resources during a training data development phase: *active learning* and *crowdsourced annotations*. Active learning refers to a process whereby the documents to be labelled are selected according to some measure which will maximize the anticipated usefulness of the training data — for example, selecting documents with diverse content (Tong and Koller, 2002).

Crowdsourcing concerns the use of annotations from non-expert annotators, relying on

identifying usable training data by using methods to assess annotator and annotation quality over a large number of low-cost annotations. One noteworthy recent work addresses the task of using multiple annotators from a machine learning perspective (Raykar et al., 2010). A typical way of deriving a gold standard of labels from a set of documents labelled by multiple annotators would be to assign the majority label to each document. Raykar et al. propose a more sophisticated model, whereby a ground truth is estimated from multiple noisy labels with the explicit intention of using these labels for training a classifier. Their approach uses the Expectation-Maximization algorithm to iteratively compute the maximum-likelihood and converge towards an optimum set of model parameters. In essence, they wish to optimize the weights they assign to the annotators, in order to best generate a gold standard and train a classifier. Their model also allows the true positive rate (specificity) and the sensitivity (1 - false positive rate) to be varied allowing the Receiver Operating Characteristic (ROC) curve to be drawn. They find the area under the ROC curve for their approach (AUC) is 3% higher than a majority voting method. This work is based on much earlier work which examined the problem of estimating annotator error-rates using the EM algorithm, but outside the context of machine learning (Dawid and Skene, 1979). This has particularly become an active research area recently with the prevalence of crowdsourced annotation tools, such as Mechanical Turk[3], where labels are low-cost, but often at the expense of quality.

Using techniques such as active learning and crowdsourced annotations likely would have improved our training data quality, however such an exercise is outside the scope of this work, and not necessary to address our research questions.

Our annotators had most trouble distinguishing between either of the sentiment-bearing classes (*positive* and *negative*) and *neutral*. This reflects the concerns raised during the training process by the annotators concerning the precise definitions of topic-directed sentiment. See Table 4.3 for a class-by-class breakdown of the doubly-annotated documents.

It should be noted that only approximately one third of the documents annotated contained sentiment, and that the ratio of sentiment-bearing documents to relevant documents which do not bear any sentiment is roughly 1:1; it is clear that separating neutral

[3] `https://www.mturk.com/`

	Microblogs	Blogs	Microreviews	Movies
Topics	Trending Topics	IR queries	Movie, App, Game, Music, Books	Films
Source	Twitter	Blogspot	Blippr	Newsgroups
Date	2009	2006	2010	pre-2003
Classes	pos/neg/neu	pos/neg/neu	pos/neg	pos/neg
Docs/class	1,000	1,000	1,000	1,000
Content	Posts (≤140 chars)	Posts	Reviews (≤140 chars)	Reviews
Annotations	Annotators	Annotators	Author	Author
Mean Words	17.885	1262.528	18.814	747.292
Mean Sentences	2.1733	72.22	1.962	32.36

Table 4.4: Sentiment corpora details

documents from documents containing sentiment is a vital part of the process. On the whole, this is encouraging for sentiment analysis however, as roughly 50% of our relevant Twitter posts contain sentiment of some kind. A situation where sentiment was more scarce would prove significantly more problematic.

4.2.2 Comparison Corpora

To contrast with our microblogs corpus, we derive a corpus of blog posts from the TREC *Blogs06* corpus (MacDonald and Ounis, 2006). We identified the most prevalant blogging platform in the corpus as Blogspot[4] (now Blogger), still one of the most commonly used blogging services. Blogger is used by a wide variety of bloggers so our data is not confined to a specific style of blog. For example, had we chosen LiveJournal[5], our data would have been biased towards journal-, or diary-style blogs. We used a templating approach to extract positive, negative and neutral blog post content and comments from the corpus, using the TREC relevance judgments as sentiment labels. Templating is a process whereby we use the `HTML` structure of the web pages to isolated the `DIV` elements which contain the blogpost text and title. This is very effective as it is common for blogs which are hosted by the same service to have a common structure. As a document may be annotated for more than one topic, documents were not used if they had been annotated with different labels for different topics. The TREC topics are diverse in nature, similar to those we used as sentiment targets in our microblog corpus.

As much of sentiment analysis literature concerns review classification, in parallel to

[4] `http://www.blogger.com`
[5] `http://www.livejournal.com`

our experiments on the microblog and blog corpora, we also conduct our experiments on a corpus of microreviews and a corpus of reviews. The reviews corpus we use as comparison is perhaps the mostly widely studied sentiment corpus, Pang and Lee's movie review corpus (Pang and Lee, 2004). This corpus contains archival movie reviews from `USENET`. In January 2010 we collected microreview documents from the microreview website, Blippr[6]. Blippr reviews bear a similarity to microblog posts in that they share the same character limit of 140 characters. Microreviews on Blippr are given one of four ratings by the author, in order from most negative to most positive: *hate*, *dislike*, *like* and *love*. In our corpus we use reviews with strongly polarised sentiment, just as they have done in constructing the movie review corpus: *hate* and *love*.

We refer to the microblog and microreview datasets as the *short-form* document corpora and the blog and movie review datasets as the *long-form* document corpora.

Our datasets are limited to exactly 1000 documents per class in line with the movie review corpus. This allows us to eliminate any underlying sentiment bias which may be learned by the classifiers. While this is obviously a consideration for a real-world system, in our experiments we wish to examine the challenges of the classification without biasing our evaluation towards the features which are discriminative for a particular class. As the sentiment distribution is different in each of the domains, this also makes accuracies comparable across datasets. We discuss the effects of uneven sentiment class distribution and our approach for dealing with classifying minority classes in Chapter 6.

4.2.3 Classification

For our experiments we use two classifiers, support vector machines (SVM) and multinomial Naive Bayes (MNB), giving us an accurate representation of the state-of-the-art in text classification. We use an SVM with a linear kernel and the cost parameter, `c`, set to 1. Optimising classifier parameters and/or using alternative kernels most likely would improve performance, however such an exercise is outside the scope of this work.

We split our corpora into sentences, and then tokenized each sentence. This was not necessary for the movie review corpus as it is distributed already split into tokenized

[6] `http://www.blippr.com`

sentences. The parts-of-speech (POS) in the data were tagged using the Stanford Part-of-Speech Tagger (Toutanova and Manning, 2000).

In our experiments, each feature in a vector records only the presence of a feature rather than the frequency of the feature in the document. This has been shown to be more effective that frequency-based feature vectors for sentiment classification (for example Pang et al. (2002)). We confirmed this on each of our datasets in preliminary experiments. We also found no benefit from stopwording or stemming. Where possible, we replaced topics with pseudo-terms to avoid learning topic-sentiment bias. We also replace URLs with a pseudo-term to avoid confusion during tokenization and POS tagging[7] Each feature vector is L2 Normalized before classification. In order to reduce computational complexity, only features which occurred four or more times in the longer corpora were used, as Pang and Lee did in their original movie review experiments. For the microblogs and microreviews datasets, all features were used as the vocabulary was much smaller. Indeed removing features renders a subset of the documents empty. We confirmed that this measure does not significantly affect classification accuracy. Accuracy was measured using 10 fold cross-validation and the folds were fixed for all experiments.

As a baseline for binary (*positive*/*negative*) classification we developed a classifier based on a sentiment lexicon, SentiWordNet (v. 1.0.1) (Esuli and Sebastiani, 2006). SentiWordNet associates positivity and negativity scores with each WordNet synset. A synset is a meaning associated with one or more word senses (Fellbaum, 1998). A word belongs to one synset for each of its senses. A synset often contains senses of more than word. In SentiWordNet for example, synset `01116026` in SentiWordNet contains senses of the words "good" and "honest" which mean "not forged; 'a good dollar bill'"[8]. This, unsupervised classifier calculates the mean positive word score and mean negative word score for a given document using the mean sentiment scores of synsets its words belong to. More formally, we consider the positive SentiWordNet score for a word w, to be the mean of the positive scores for all the synsets of that word which have the same WordNet POS:

[7]The pseudo-term used to replace the topics and users were *UserString* and *TopicString*. Due to the -ing ending, these were mistakenly tagged as verbs by the POS tagger, where they would be better tagged as nouns. This impacted the POS feature vectors, though it is unlikely it had a significant impact on the results.

[8]An online searchable version of SentiWordNet is available at `http://sentiwordnet.isti.cnr.it`.

$$s_{pos}(w) = \frac{1}{n}\sum_{i=0}^{n}\left(\frac{1}{m}\sum_{k=0}^{m} PosSwn_{i,k}\right) \tag{4.1}$$

where n is the number of synsets the word appears in, m is the number of word senses in the synset for that word and $PosSwn_{i,k}$ is the positivity score for word sense k in synset i for word w. WordNet POS for a synset is one of noun, verb, adjective, adverb, and represents each of the senses in the synset. The senses in a synset typically have the same POS. By mapping our POS tags from the Stanford POS tagger to the WordNet POS, we achieve a degree of word sense disambiguation.

The positive score for a document is the mean $s_{pos}(w)$ for all words in the document and is given by:

$$score_{positive}(d) = \frac{1}{p}\sum_{i=0}^{p} s_{pos}(w_i) \tag{4.2}$$

for a document d with p words. The negative score is calculated similarly. If $score_{negative}(d)$ is greater than $score_{positive}(d)$, d is classified as *negative*; otherwise d is classified as *positive*. If the scores are equal, an arbitrary class is assigned.

In all cases, words are stemmed using a WordNet stemmer[9], and only word senses with a matching POS are considered. We have used this approach with success in earlier works (Bermingham et al., 2008, 2009). We use this classifier to demonstrate the ability of a relatively trivial unsupervised classifier, in contrast to a supervised classifier. Despite their naivety, this type of classifier is often used as it does not require expensive training data.

There are two extensions to this classifier which likely would have proved beneficial to the classifier's accuracy. The first is term negation, whereby token polarity may be reversed if associated with a negating token such as "not" or "no". Another technique which could have helped, is incorporating topic-token proximity giving a relevance weighting to sentiment values.

[9] `http://www.rednoise.org/rita/wordnet`

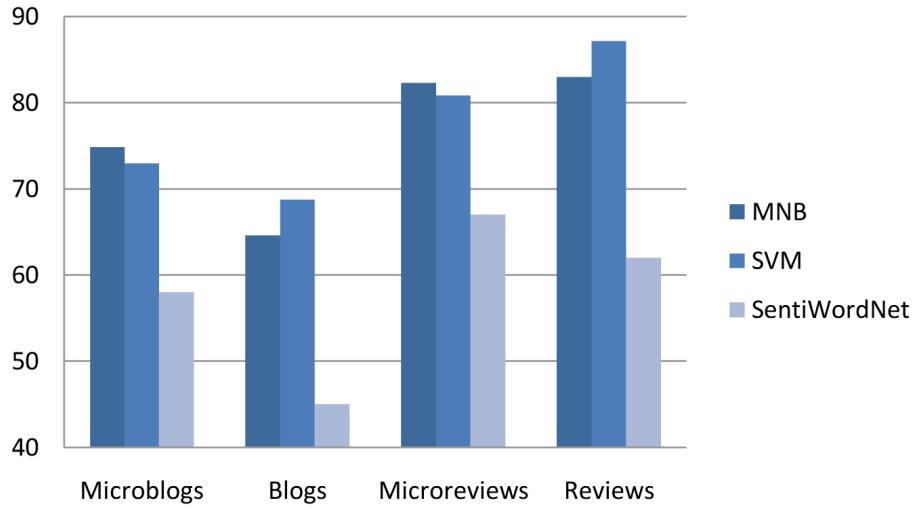


Figure 4.2: Percentage sentiment classification accuracies for unigram features

4.3 Results and Discussion

The results are shown in Table 4.5 with a comparison of binary unigram classification accuracies across collections in Figure 4.2. Unigram binary classification accuracy for microblogs is 74.85% using an SVM. This is an encouraging accuracy given the diversity in the sentiment topics. As we have balanced datasets, a classifier which assigns labels randomly would achieve approximately 50% accuracy for binary classification. For microreviews, the accuracy is considerably higher than for microblogs, at 82.25% using an SVM. As expected, the classifier finds it easier to distinguish between polarised reviews than to identify sentiment in arbitrary posts.

Sentiment classification of the long-form documents yields some surprising results. Blog classification accuracy is significantly lower than for microblogs. However, movie review classification is higher than for microreviews, confirming Pang and Lee's result of 87.15% for SVM with unigram features. At first this may seem contradictory — surely the classifier should perform consistently across textual domains? We speculate that this behaviour is due to within-document topic drift. In the two review corpora the text of the document has a high density of sentiment information about the topic, and a low noise density. In the blogs dataset, this is not necessarily the case; the sentiment in a blog post may be an isolated reference in a subsection of the document. One approach to mitigate this affect is to create metadocuments consisting of topically relevant subsections and use

Feature Set	Microblogs		Blogs		Microreviews		Movies	
	MNB	SVM	MNB	SVM	MNB	SVM	MNB	SVM
Unigram	**74.85**	72.95	64.6	68.75	**82.25**	80.8	82.95	87.15
+Bigram	74.35	72.95	64.6	68.45	82.15	**81.4**	85.25	**87.9**
+Bigram+Trigram	73.7	72.8	64.6	68.5	81.95	80.85	84.8	**87.9**
+POS n-gram (n=1)	73.25	71.6	64.7	68.45	80.8	79.5	82.4	86.95
+POS n-gram (n=1,2)	70.25	70.05	62.6	66.25	80.8	79.5	81.8	84.95
+POS n-gram (n=1,2,3)	68.8	69.7	62.45	64.6	74.7	76.9	79.95	82
+POS-STW Bigram	74.15	73.25	64.5	**69**	82.5	81.05	85.35	87.5
+POS-STW Bigram+Trigram	74.4	**73.45**	**64.85**	68.7	82.15	80.6	**85.5**	87.8

Table 4.5: Percentage accuracy for binary classification

these as training and test documents. We have employed this with success in previous blog sentiment classification experiments (O'Hare et al., 2009). Topic drift also occurs in the microblog corpus; indeed this fact was reported by our annotators. However, given the shorter documents, there is less opportunity for noisy, non-relevant information to enter the feature vector and our classifier is not as adversely affected as in the blog domain.

Our unsupervised, lexicon-based classifier performs poorly across all datasets. For the blogs corpus, the accuracy is less than 50%. The accuracy gap between supervised and unsupervised classification accuracy in the long-form corpora is much more pronounced. This makes intuitive sense as the probability of the polarity of a given word in a document expressing sentiment towards a topic is again much higher in the short-form domains.

Of the two supervised classifiers, SVM outperforms MNB in the long-form domains, whereas the opposite is true in the short-form domains. SVMs scale better with larger vector dimensionality (Joachims, 1998) so this is most likely the reason for this observation; the number of unique terms in the longer documents is over three times their shorter counterparts, even when infrequent features have been excluded.

Having established a reasonable performance in sentiment classification of microblog posts, we wish to explore whether we can improve the standard bag of words feature set by adding more sophisticated features. Using sequences of terms, or n-grams, we can capture some of the information lost in the bag-of-words model. We evaluated two feature sets: (unigrams + bigrams) and (unigrams + bigrams + trigrams). We found that although an increase in classification accuracy is observed for the movie reviews, this is not the case for any of the other datasets (see Table 4.5). We also examined POS-based n-grams in conjunction with a unigram model and observed a decrease in accuracy across all corpora.

POS tag	example	POS tag	example
AUX	do done have is	NNPS	Americans Amharas
CC	and both but either	PDT	all both half many
CD	one-tenth ten million 0.5	POS	' 's
DT	all an the them these this	PRP	hers herself him himself
EX	there	PRP$	her his mine my
FW	gemeinschaft hund ich jeux	RP	aboard about across along
IN	astride among uppon whether out	SYM	% & ' "
LS	SP-44005 SP-44007 Second Third	TO	to
NNP	Motown Venneboerger Ranzer	WDT	that what whatever which

Table 4.6: POS tags stopworded using Matsumoto technique for removing common POS tags from n-grams. Table from Matsumoto et al. (2005).

This indicates that the syntactic patterns represented by the POS n-gram features do not contain information which is more discriminative than unigrams. It should be noted that the POS tagger we used has not been trained on well formed text, and thus likely has a lower accuracy on our short-form domains.

The most promising results came from a POS-based stopwording approach proposed by Matsumoto *et al.* (Matsumoto et al., 2005) (see Figure 4.6). This approach (which Matsumoto *et al.* refer to as "word sub-sequences") consists of an n-gram model, where terms have been stopworded based on their POS. We use the same POS list as Matsumoto. These features increase accuracy across all corpora for unigrams + POS-stopworded bigrams. This suggests that a better understanding of the linguistic context of terms is similarly advantageous in all domains.

Examining the discriminative features across the datasets gives us a unique insight into the important features for sentiment classification. We use a standard measure of discriminability, information gain ratio. This measure is particularly useful as it does not favour features which occur frequently in the training set. The 25 most discriminative unigrams, bigrams and trigrams for binary classification in each dataset are listed in Table 4.7. Immediately obvious is the significant role that punctuation plays in expressing sentiment in microblog posts. Emoticons, exclamation marks, quotation marks, questions and ellipses are all among the most discriminative features for microblogs, yet they do not rank highly among the most discriminative features in the other datasets. This suggests that these are being used specifically in microblog posts to express sentiment, perhaps

	Microblogs	Blogs	Microreviews	Reviews
1	!	witherspoon	great	bad
2	<Urlstring>	joaquin	boring	worst
3	<Topicstring>	reese witherspoon	best	stupid
4	amazing	joaquin phoenix	terrible	boring
5	. .	sharon	the best	the worst
6	! !	ledger	worst	waste
7	?	heath ledger	n't	ridiculous
8	! ! !	heath	love	wasted
9	love	johnny cash	loved	awful
10	<Topicstring> !	palestinians	?	?
11	great	philip	the worst	outstanding
12	bonuses	gyllenhaal	awesome	mess
13	not	greenhouse	amazing	supposed
14	by	iranian	did	life
15	awesome	seymour	did n't	lame
16	win	jerusalem	boring	have
17	:)	doctors	classic	waste of
18	i	prejudice	great movie	nothing
19	:	june carter	bad	of the best
20	see	and watch	crap	dull
21	happy	cartoons of	one of	best
22	i love	jake gyllenhaal	waste	supposed to
23	on	favourite	love it	should have
24	‘ ‘	seymour hoffman	of the best	plot
25	forward	lobbying	not	unfortunately

Table 4.7: Most discriminative unigram, bigram and trigram features for each dataset according to information gain ratio for binary classification

	MNB	SVM	#features
Microblogs	61.3	59.5	8132
Blogs	52.13	57.6	28805

Table 4.8: Three-way unigram sentiment classification percentage accuracies: *positive*, *negative*, *neutral*

as indicators of intonation. Identifying precisely how these features are being used in microblogs remains an exercise for future work, though they do provide an opportunity to engineer features which can capture these domain-specific artefacts. The fact that is not observed for microreviews is interesting and is possibly an artefact of the difference in modalities through which the content was created, or perhaps simply reflective of a deeper distinction in the nature of the content.

The discriminative features for both the reviews and microreviews are largely similar in nature, typically polarised adjectives. The blog classifier appears to have learned a certain amount of entity bias, as many of the discriminative features are people or places. Note that none of these entities are topic terms (topic terms were removed in pre-processing), though they do appear to be entities associated with topics. With the classifier overfitting to these terms, it is clear that the blog classifier had trouble identifying generic discriminative features.

With any discussion on sentiment analysis in non-review domains, it is important to note that there will always be *neutral* documents. As we saw in Section 4.2, for every positive or negative document, there was one neutral document annotated. Assuming that we can identify relevance in documents, we must still separate the sentiment-bearing documents, from the non-sentiment bearing. Generally, if computational resources are not a concern, a three-way classifier, which classifies documents as either *positive*, *negative* or *neutral* is sufficient. Results of our three-way classification on microblogs and blogs can be seen in Table 4.8. The accuracy is, as expected, significantly less than for binary classification with SVMs again outperforming MNB on the longer blog documents, though note that in this case, the accuracy of a classifier which assigns labels randomly is approximately 33.33%.

4.4 Conclusion

In this chapter, we have explored sentiment analysis in microblog posts using supervised learning. We used multinomial Naive Bayes and support vector machine classifiers, as well as an unsupervised, lexicon-based classifier. We evaluated a range of feature sets, including n-grams and POS-based feature sets. Our focus has been to identify what, if any, unique challenges exist in classifying such short documents. To accomplish this, we have contrasted our results with another short-from textual domain, microreviews, as well as two long-form document collections, blogs and movie reviews. We also examined three-way classification, taking into account the neutral annotations in our microblog and blog corpora. Finally, we used a discriminability measure, information gain ratio, to explore the relative significance of features in the various corpora.

The results of our experiments on the whole are encouraging for the task of analysing sentiment in microblogs. We achieve an accuracy of 74.85% for binary classification for a diverse set of topics, indicating we can classify microblog documents with a moderate degree of confidence. In both of our short-form corpora, we find it difficult to improve performance by extending a unigram feature representation. This is contrary to the long-form corpora which respond favourably to enriched feature representations. We do however see promise in sophisticated POS-based features across all datasets and speculate that engineering features based on deeper linguistic representations, such as syntactic parse trees in the form of dependency and phrase structure analyses, may work for microblogs as they have been shown to do for movie reviews.

We find that supervised classification performs far better than an unsupervised, lexicon-based classifier, and that this effect is more pronounced in the long-form corpora. We also find that MNB outperforms SVMs for classifying the short form documents, whereas the opposite is true for the long-form documents.

In analysing discriminative features, we find that a significant role is played by punctuation in expressing sentiment in microblog posts. This is in line with microblogs as an informal CMC domain, containing elements of speech-like text. It is surprising to see that this is not a pattern seen in our microreviews corpus, indicating that this is not an artefact of all short-form platforms.

On the whole, we see commonalities between the two short-form corpora, and between the two long-form corpora. We also see commonalities between the two review corpora, and between the two blog corpora. We conclude that although the shortness of the documents has a bearing on which feature sets and classifier will provide optimum performance, the low number of features present in the documents does not hamper sentiment classification. On the contrary, we find classifying these short documents a much easier task than their longer counterparts, blogs. Also, the "noisy" artefacts of the microblog domain, such as informal punctuation, turn out to be discriminative. These results provide a compelling argument to encourage the research community to focus on microblogs in sentiment analysis research.

We conclude from these results that sentiment analysis in microblogs using supervised machine learning is suitable for use in our search experiments. At 75% for binary sentiment classification and 61.3% for three-way sentiment classification accuracy, our classifiers have demonstrated considerable ability to discriminate between documents with respect to their sentiment. Furthermore, this performance is attainable with a unigram approach, and is not contingent on any complex, resource-intensive feature extraction; it is therefore pragmatic for use in a real-time system. As we prepare more focused topics and training data for the real-time Channel S experiments, we can expect our accuracy to improve over the generic classifier evaluated in this chapter. It should be noted however, that at this level of performance, the classifier is still making a significant number of misclassifications. It is possible that this could mean we do not observe a sentiment-related effect which is present in our real-time experiments, due to the noise in the classifier output.

Since the completion of these experiments, there have been further works which have applied sentiment analysis to microblogs. O'Connor et al. observe leading signals in Twitter sentiment with respect to political opinion polls (O'Connor et al., 2010a). Others have explored the potential of tracking sentiment to predict movie sales (Asur and Huberman, 2010), election results (Tumasjan et al., 2010) or the Dow Jones Industrial Average (Bollen et al., 2011). The diversity of these studies illustrate the potential range of applications for microblog sentiment analysis, particularly with respect to approximating or predicting real-world values. They confirm our conclusion that sentiment analysis in microblogs is feasible and suitable to support a variety of applications.

Chapter 5

Simulated Real-time Evaluation

5.1 Introduction

In the previous chapter we saw how microblog documents, or posts, may be classified according to sentiment to a significant degree of accuracy (74.85% for binary classification). We now wish to progress towards our goal of building our classifiers into a real-time scenario. However, before deploying a real-time system, we want to learn more about the dynamics of real-time search scenarios and, in particular, the role played by sentiment. To address this, we devise an experiment which uses our manually labelled microblog sentiment data to create simulated real-time search scenarios, and conduct a user study evaluation. We make a number of observations relating to sentiment with respect to the participants, topics, the documents themselves, and a number of sentiment-based filtering algorithms.

There is an inherent immediacy with real-time scenarios. Often topics of real-time interest, such as breaking news stories, cannot be identified in advance. In other cases, prescient knowledge is available, for example for scheduled sports events or television programmes. For this reason, we break our real-time evaluation into two stages. In this chapter, we *simulate* real-time scenarios so that we can examine a diverse range of topics. Then, in Chapter 6, we pursue two topics at a much deeper level, with live, real-time user studies.

The evaluation of real-time systems is a troublesome proposition. On one hand, with

the benefit of hindsight, it can be easier to see at a point in time in the past what the valuable information and commentary had been. On the other hand, hindsight does not account for the real-time user experience, and the specific nature of the real-time information need. It is this real-time user task and feedback that form the focus of our evaluations. In soliciting feedback from users in real-time, we can capture their immediate appreciation and dislike of different types of information in the stream.

In simulating the real-time scenario in the experiment presented in this chapter, we run the risk of participants' *a posteriori* knowledge effecting their perception of the information. However, the interface we present to the user is as close to a realistic real-time environment as is possible, and the concessions we make through simulation are compensated for in our ability to assess a variety of real-time topics in a laboratory setting. We can also use document-topic pairs which are manually labelled for sentiment, giving us an analog for a high precision sentiment classifier.

In this chapter we first give an overview of the methodology in Section 5.2. This is followed by our experimental results in Section 5.3 and discussion in Section 5.4. We conclude in Section 5.5.

5.2 Methodology

Recall our research questions:

- *Do sentiment-based algorithms differ significantly from a baseline sampling approach?*
- *Do users' demographics and preferences significantly affect their perception of sentiment? Which types of sentiment have the most profound impact?*
- *Is sentiment a predictor of whether individual documents will be regarded as important by users?*

In the following sections we look at the aspects of the experimental set-up that we use to address these questions: the topics, the experimental design, and our methods of measurement and evaluation.

5.2.1 Topics

In Chapter 4, we used 50 topics to conduct our supervised learning evaluation. For the purpose of this next experiment however, this number of topics is to large, and we need to select a subset to use in search scenarios. In order to facilitate comparison across topics, we ensure that topics are distributed evenly across users and algorithms. In this section we take a closer look at the topics, and the sentiment annotations for their documents.

Each sentiment annotation associates a label with a $< Topic, Document >$ pair. Note that a document may be relevant for more than one topic. First, we disregard any annotations which were labelled *unannotatable* or *unclear*; these labels do not carry any sentiment or relevance information. A portion of the annotations were selected for testing inter-annotator agreement and thus have more than one label. If a $< Topic, Document >$ pair had multiple conflicting labels, we discarded all labels for that pair. This gives us a set of $< Topic, Document >$ pairs, each with a single label: *positive*, *negative*, *neutral*, *mixed* or *not relevant*, amounting to an average of 107 labelled documents for each topic.

There were three topics for which more than 50% of the documents were annotated non-relevant: *"Fargo"*, *"budget"* and *"Wales"*. The high degree of non-relevant documents for these topics proved to be due to topic ambiguity. The topic, *budget*, referred to the United States Federal Budget — it was the announcement of the budget that caused this topic to trend. However, "budget", is a common term in our corpus and many of the documents presented to annotators contained other uses of the term "budget", or references to the budgets of other countries. Similarly, there were ambiguities for the topic *Wales* (the rugby team or the country) and Fargo (the film or Wells Fargo, the financial services company). Aside from this, 80% of the topics had fewer than 10% non-relevant documents. 11 topics had no non-relevant documents. This is encouraging for our naive relevance measure, which considers a document relevant if it contains the topic terms. Relevance precision could likely be improved for topics with low relevance by simply introducing disambiguating terms into the topic query, assuming the consequent reduction in recall is acceptable. We also likely would have been able to increase relevance precision if we used data only within topic-specific time bounds. A recent approach in the literature uses bootstrapping with known relevant documents to classify relevant microblog posts

for filtering microblog streams for television programme topics (Dan et al., 2011).

We define two metrics to represent the distribution of sentiment in the labelled documents for each topic so that we may represent the sentiment bias in our labelled data. Firstly, we define the subjectivity for topic t to be the proportion of the relevant documents which contain sentiment of any kind:

$$Subj(t) = \frac{|d_{t,pos}| + |d_{t,mix}| + |d_{t,neg}| - |d_{t,neu}|}{|d_{t,pos}| + |d_{t,mix}| + |d_{t,neg}| + |d_{t,neu}|} \tag{5.1}$$

where $|d_{t,x}|$ is the number of documents relevant to topic t with the label x. Similarly, we define sentiment for topic t as the proportion of positive documents minus the proportion of negative documents:

$$Sent(t) = \frac{|d_{t,pos}| - |d_{t,neg}|}{|d_{t,pos}| + |d_{t,mix}| + |d_{t,neg}| + |d_{t,neu}|} \tag{5.2}$$

Using these two measures, we visualise the topics in Figure 5.1. We observe a significant positive correlation between the level of subjectivity expressed for a given topic and the sentiment for that topic ($r = 0.43$, $p < 0.001$); the higher the proportion of sentiment that is expressed about a topic, the more likely that the net sentiment will be positive. If we decompose this set of topics into the five topic categories, we can see that for four of the topic types (*Entertainment*, *Sports*, *Politics and Government*, *Products and Services*) the correlation is positive but for one category, *Companies*, the correlation is negative (see Table 5.2). On further examination of the graph for *Politics and Government*, we see that there are two outliers without which this category would exhibit a similar pattern to *Companies*.

For the simulated search scenario we chose topics which (i) had a low proportion of non-relevant documents and (ii) were real-time in nature - typically unfolding news stories or live events and (iii) which were familiar to our users. We also ensured a coverage across our topic categories. The chosen topics are **bolded** in Table 5.1.

5.2.2 Experimental Set-up

We recruited 16 participants for our study. They consisted of faculty staff and postgraduate students who volunteered to take part. Our document annotators were not permitted

ID	Topic	pos	neg	neu	not rel	mix	total rel	not rel %	Sent(t)	Subj(t)
1	**Susan Boyle**	70	15	30	0	4	119	0.00	0.46	0.50
2	**Twilight**	58	17	27	19	4	106	15.20	0.39	0.49
3	Leno	38	14	68	1	2	122	0.81	0.20	-0.11
4	**Bono**	19	30	19	15	4	72	17.24	-0.15	0.47
5	Adam Lambert	90	15	21	1	4	130	0.76	0.58	0.68
6	**Watchmen**	57	15	38	3	9	119	2.46	0.35	0.36
7	Rihanna	13	23	76	0	4	116	0.00	-0.09	-0.31
8	Fargo	3	0	4	125	2	9	93.28	0.33	0.11
9	**Red Dwarf**	69	25	30	0	10	134	0.00	0.33	0.55
10	Coachella	63	7	48	1	6	124	0.80	0.45	0.23
11	**Man Utd**	41	29	59	0	3	132	0.00	0.09	0.11
12	**Celtics**	51	26	48	2	2	127	1.55	0.20	0.24
13	**Arsenal**	36	16	45	8	1	98	7.55	0.20	0.08
14	Tiger Woods	81	5	28	2	1	115	1.71	0.66	0.51
15	**Lance Armstrong**	31	12	63	2	4	110	1.79	0.17	-0.15
16	Curt Schilling	41	15	65	1	6	127	0.78	0.20	-0.02
17	**Mets**	42	19	52	5	5	118	4.07	0.19	0.12
18	Buffalo Bills	12	24	80	3	4	120	2.44	-0.10	-0.33
19	Terrell Owens	17	31	64	0	2	114	0.00	-0.12	-0.12
20	Wales	17	5	11	98	2	35	73.68	0.34	0.37
21	North Korea	2	43	88	1	0	133	0.75	-0.31	-0.32
22	NATO	5	7	65	0	1	78	0.00	-0.03	-0.67
23	Afghanistan War	7	35	78	4	0	120	3.23	-0.23	-0.30
24	Dave Ramsey	56	6	59	3	1	122	2.40	0.41	0.03
25	Rush Limbaugh	7	76	37	0	1	121	0.00	-0.57	0.39
26	Navy SEALS	71	5	38	9	1	115	7.26	0.57	0.34
27	Gordon Brown	5	49	68	0	2	124	0.00	-0.35	-0.10
28	Sanjay Gupta	11	23	93	2	0	127	1.55	-0.09	-0.46
29	**Obama**	16	36	58	4	1	111	3.48	-0.18	-0.05
30	budget	4	14	13	88	1	32	73.33	-0.31	0.19
31	Kindle	49	25	39	1	7	120	0.83	0.20	0.35
32	Wolfram Alpha	46	7	54	3	7	114	2.56	0.34	0.05
33	Guinness	73	6	36	11	2	117	8.59	0.57	0.38
34	Pirate Bay	11	13	73	3	4	101	2.88	-0.02	-0.45
35	Skype	30	6	68	1	1	105	0.94	0.23	-0.30
36	Sky News	5	25	72	29	2	104	21.80	-0.19	-0.38
37	Nikon D5000	23	9	71	0	2	105	0.00	0.13	-0.35
38	**Safari 4**	49	31	22	3	16	118	2.48	0.15	0.63
39	**iPhone**	25	12	70	2	4	111	1.77	0.12	-0.26
40	Spotify	36	9	40	12	1	86	12.24	0.31	0.07
41	AIG	1	71	45	1	1	118	0.84	-0.59	0.24
42	Oracle	4	13	65	16	0	82	16.33	-0.11	-0.59
43	**Wal-Mart**	19	35	67	0	2	123	0.00	-0.13	-0.09
44	Sun Microsystems	3	12	88	0	2	105	0.00	-0.09	-0.68
45	CNBC	14	39	63	19	1	117	13.97	-0.21	-0.08
46	Chrysler	6	31	63	2	1	101	1.94	-0.25	-0.25
47	Lloyds	7	35	63	7	0	105	6.25	-0.27	-0.20
48	**IBM**	13	19	72	2	4	108	1.82	-0.06	-0.33
49	Toyota	16	10	81	2	3	110	1.79	0.05	-0.47
50	ACMA	5	39	66	14	4	114	10.94	-0.30	-0.16
	mean	29.36	21.68	53.82	10.5	3.02	107.88	8.87	0.07	0

Table 5.1: Topic annotation counts and subjectivity and sentiment scores (Topics used in simulated evaluation in **bold**)

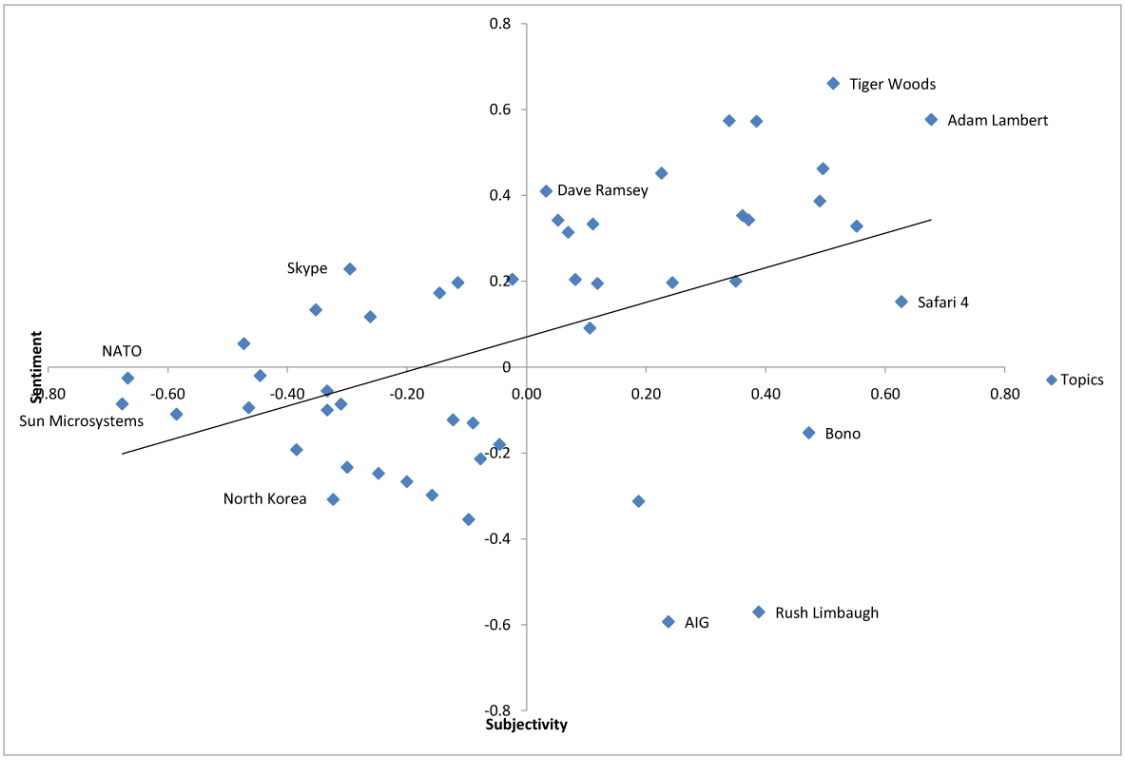


Figure 5.1: In our labelled documents, topic subjectivity was positively correlated with topic sentiment ($r = 0.48$, 2-tailed, $p < 0.001$)

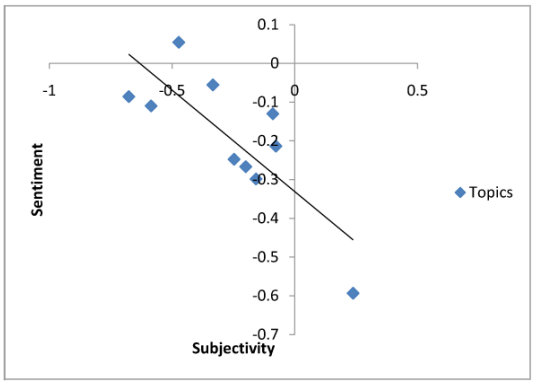
to participate as they were already familiar with the task and topics.

We instructed the participants that they were testing a new system for monitoring Twitter during live events. They were not made aware that the focus of our evaluation was sentiment. As discussed in Chapter 3, there are two types of feedback we require participants to give to the system:

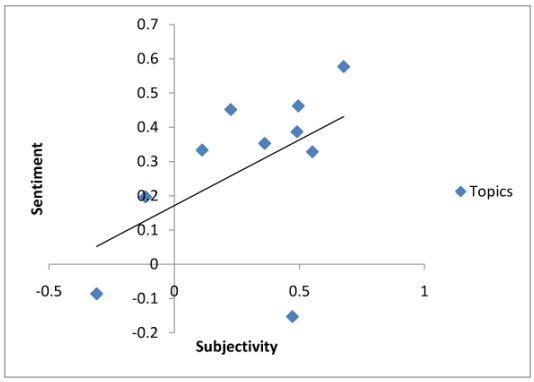
- *document-level*: User may *thumbs up* or *thumbs down* a document as it appears in a stream.
- *stream-level*: At the end of a stream, a user must rate the stream for how *insightful*, *informative* and *interesting* they found it, and then give it an overall rating.

We also captured some profile information about the participants and recorded their prior sentiment towards the topics. See Table 5.2 for sample sizes for profile attributes.

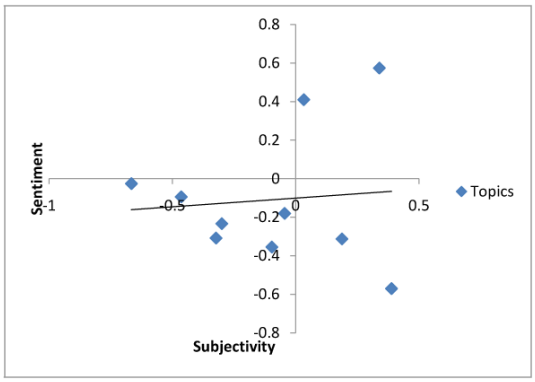
The system presents the topics to the users in batches of 12 documents (a *stream*) from our labelled set. In each stream the documents are ordered chronologically according to their original timestamp to ensure the stream is as organic as possible. We conducted



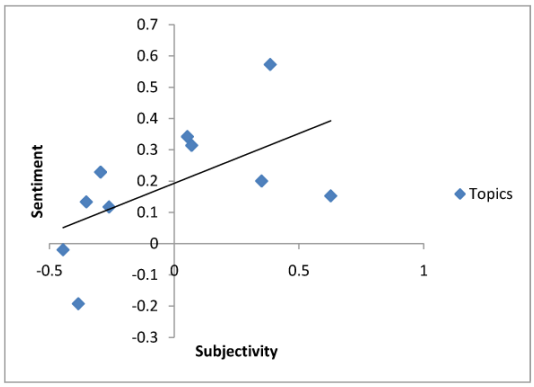
(a) Companies



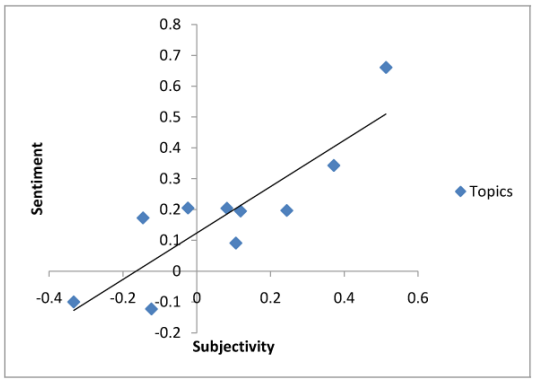
(b) Entertainment



(c) Politics and Government



(d) Products and Services



(e) Sports

Figure 5.2: Subjectivity-sentiment relationships for each topic category

Participant Profile		count
Age	≥25	11
	<25	5
Task Familiarity	unfamiliar	9
	slightly or more familiar	7
Gender	female	3
	male	13

Table 5.2: Participant sample sizes for profile attributes

a pilot test with two non-participants and they deemed an interval of 10 seconds between documents appearing in a stream to be comfortable for the task.

As we know the sentiment of each document in advance, this means we can employ algorithms for streams that filter certain types of sentiment. To this end, we devised eight algorithms which selected documents for a topic stream according to their sentiment:

- `pos`: 12 positive documents
- `neg`: 12 negative documents
- `neu`: 12 neutral documents
- `posneg`: 6 positive documents, 6 negative documents
- `posneu`: 6 positive documents, 6 neutral documents
- `negneu`: 6 negative documents, 6 neutral documents
- `posnegneu`: 4 positive documents, 4 negative documents, 4 neutral documents
- `control`: 12 positive, negative or neutral documents randomly sampled from the annotations

These algorithms were assigned to the $< User, Topic >$ pairs in a Latin squares arrangement (see Table 5.3). This ensured that each user encountered each algorithm twice, and that the algorithms were evenly distributed throughout the topics. The order of the topics was randomized. See Table 5.4 for the final ordering of $< Topic, Algorithm >$ pairs for each user.

We instructed participants to thumbs up documents if they would like the system to show more similar documents and thumbs down documents they would rather the system did not present to them. They were told they were under no obligation to give explicit document-level feedback to the system, so some documents they could simply leave with no feedback. After each stream, the system asked the user to fill out a short survey. They could adjourn the experiment between topic streams and resume at a later time, but if they had started a topic, they must complete it before taking a break. The majority of the participants chose to complete the experiment in one session. The system presented

	1	2	4	6	9	11	12	13	15	17	29	31	38	39	43	48
0	**1**	**1**	2	2	3	3	4	4	5	5	6	6	7	7	8	8
1	8	8	**1**	**1**	2	2	3	3	4	4	5	5	6	6	7	7
2	7	7	8	8	**1**	**1**	2	2	3	3	4	4	5	5	6	6
3	6	6	7	7	8	8	**1**	**1**	2	2	3	3	4	4	5	5
4	5	5	6	6	7	7	8	8	**1**	**1**	2	2	3	3	4	4
5	4	4	5	5	6	6	7	7	8	8	**1**	**1**	2	2	3	3
6	3	3	4	4	5	5	6	6	7	7	8	8	**1**	**1**	2	2
7	2	2	3	3	4	4	5	5	6	6	7	7	8	8	**1**	**1**
8	**1**	**1**	2	2	3	3	4	4	5	5	6	6	7	7	8	8
9	8	8	**1**	**1**	2	2	3	3	4	4	5	5	6	6	7	7
10	7	7	8	8	**1**	**1**	2	2	3	3	4	4	5	5	6	6
11	6	6	7	7	8	8	**1**	**1**	2	2	3	3	4	4	5	5
12	5	5	6	6	7	7	8	8	**1**	**1**	2	2	3	3	4	4
13	4	4	5	5	6	6	7	7	8	8	**1**	**1**	2	2	3	3
14	3	3	4	4	5	5	6	6	7	7	8	8	**1**	**1**	2	2
15	2	2	3	3	4	4	5	5	6	6	7	7	8	8	**1**	**1**

Table 5.3: Algorithms were assigned to topics (columns) and users (rows) in a Latin squares arrangement: `pos=1`, `neg=2`, `neu=3`, `posneg=4`, `posneu=5`, `negneu=6`, `posnegneu=7`, `ctrl=8` (`pos` algorithm in **bold**)

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
0	31,6	38,7	15,5	6,2	12,4	17,5	39,7	9,3	**1,1**	43,8	11,3	**2,1**	13,4	29,6	48,8	4,2
1	1,8	17,4	12,3	31,5	**4,1**	38,6	39,6	9,2	**6,1**	11,2	29,5	15,4	2,8	13,3	48,7	43,7
2	43,6	6,8	**11,1**	39,5	13,2	1,7	17,3	12,2	48,6	31,4	29,4	38,5	2,7	15,3	4,8	**9,1**
3	2,6	39,4	15,2	**12,1**	1,6	6,7	38,4	11,8	9,8	48,5	31,3	17,2	43,5	29,3	4,7	**13,1**
4	**15,1**	43,4	29,2	13,8	31,2	2,5	12,8	48,4	1,5	**17,1**	6,6	38,3	4,6	39,3	9,7	11,7
5	38,2	**29,1**	9,6	43,3	17,8	39,2	**31,1**	12,7	13,7	2,4	48,3	6,5	4,5	1,4	11,6	15,8
6	6,4	15,7	11,5	48,2	29,8	2,3	13,6	12,6	1,3	43,2	31,8	**39,1**	4,4	**38,1**	17,7	9,5
7	1,2	39,8	11,4	6,3	29,7	13,5	**48,1**	12,5	**43,1**	17,6	38,8	2,2	9,4	31,7	15,6	4,3
8	11,3	**1,1**	31,6	48,8	29,6	39,7	**2,1**	4,2	12,4	9,3	15,5	6,2	13,4	17,5	43,8	38,7
9	**6,1**	15,4	17,4	48,7	**4,1**	43,7	29,5	13,3	39,6	2,8	38,6	11,2	31,5	9,2	1,8	12,3
10	38,5	**11,1**	17,3	15,3	31,4	1,7	12,2	43,6	2,7	13,2	4,8	39,5	**9,1**	6,8	48,6	29,4
11	9,8	4,7	43,5	15,2	6,7	29,3	31,3	**12,1**	17,2	38,4	48,5	1,6	39,4	**13,1**	2,6	11,8
12	9,7	2,5	38,3	**17,1**	29,2	1,5	4,6	43,4	13,8	31,2	11,7	39,3	6,6	**15,1**	12,8	48,4
13	**29,1**	38,2	13,7	4,5	43,3	48,3	17,8	2,4	15,8	1,4	12,7	11,6	**31,1**	6,5	39,2	9,6
14	1,3	9,5	6,4	13,6	29,8	11,5	17,7	**38,1**	48,2	2,3	31,8	15,7	**39,1**	4,4	12,6	43,2
15	**48,1**	38,8	29,7	9,4	6,3	13,5	15,6	1,2	39,8	2,2	17,6	31,7	11,4	12,5	4,3	**43,1**

Table 5.4: Final $< Topic, Algorithm >$ pairs for users (rows) in order (columns) after assigning a random ordering (`pos` algorithm in **bold**)

each participant with a training topic before commencing the experiment. Participants confirmed they were comfortable with the task before proceeding. Feedback from training topics was not used in our evaluation.

5.2.3 Measurement

We gather four types of data in this experiment:

- *Document-level feedback*: one of {*thumbs up*, *thumbs down*, *no feedback*} for each $< User, Topic, Document >$ combination selected by the system.
- *Stream-level feedback*: survey feedback for each $< User, Topic >$ pair in the experiment:
  - 5-point Likert scale for *interesting*, *insightful*, *informative* {*strongly disagree*, *disagree*, *neither*, *agree*, *strongly agree*}; For simplicity, we map these to 3 categories, conflating *strongly agree* and *agree*, and conflating *strongly disagree* and *disagree*.
  - 7-point overall rating for each $< User, Topic >$ from poor (1) to excellent (7).

  Where neccessary, we mapped the Likert scale to a 5-point numerical scale (e.g. to facilitate correlation). Conversely, we mapped the 7-point overall scale to three categories where needed: Poor {0,1,2}, OK {3} and Good {4,5,6}.
- *Topic distribution*: $Subj(t)$ and $Sent(t)$ for topic t distributions in annotation samples (as described in Section 5.2.1)
- *User profile*: Demographic participant information which we map to binary categories — age {*under 25*, *25 and older*}, gender {*male*, *female*} and familiarity with Twitter search {*unfamiliar* (never use Twitter search), *familiar* (use Twitter search sometimes or more often)}[1]. We also record user sentiment towards topics as one of {*unfamiliar*, *neutral*, *positive*, *negative*}.

[1]We also recorded education level, but this is not used in our evaluation, due to lack of diversity of participants.

	α
Overall	0.0392
Informative	0.1232
Interesting	0.0051
Insightful	0.0501

Table 5.5: Krippendorff's alpha reliability estimate for participant $< Topic, Algorithm >$ stream ratings

See Appendix C for all materials associated with this experiment. For a given set of documents, D, we quantify the feedback from document-level feedback as the proportion of documents which receive a thumbs up annotation minus the proportion that receive a thumbs down annotation:

$$NetFeedback(D) = \frac{|D_{thumbsup}| - |D_{thumbsdown}|}{|D|} \tag{5.3}$$

When aggregating across streams (for example for all streams where the algorithm `pos` was used), we simply use the mean of the values for all the relevant streams, both for document-level feedback, and for survey feedback.

5.3 Results

In total, we gathered feedback for 3,072 documents, across 256 streams, from 16 users. Each of the 128 unique $< Topic, Algorithm >$ combinations was presented to two participants. In this section we present our results, observations and analysis.

5.3.1 Feedback

If we look at the agreement between pairs of participants who were presented the same $< Topic, Algorithm >$ configuration, we observe almost no agreement (see Table 5.5). For these figures we use Krippendorff's alpha reliability estimate, α (Hayes and Krippendorff, 2007). α is devised to handle categorical data. so we use 3-class categorical versions of our survey scales as described in Section 5.2.3.

In terms of document-level feedback, we calculate the agreement between participants for $< Document, Topic >$ pairs (see Table 5.6). We observe a higher agreement than for

	α
Thumbs up/Thumbs down/No annotation	0.1424
Thumbs up only	0.2211
Thumbs down only	0.2159
No annotation only	0.0124

Table 5.6: Krippendorff's alpha reliability estimate for document-level feedback

	Net Feedback	Informative	Interesting	Insightful	Overall
Net Feedback	1	-0.0499	0.0068	0.0055	-0.0336
Informative		1	0.6885**	0.7153**	0.7989**
Interesting			1	0.7211**	0.68**
Insightful				1	0.7044**
Overall					1

Table 5.7: Pearson product-moment correlation coefficient between feedback measures (** indicates significant (2-sided, $p < 0.001$))

surveys but still overall quite a low level agreement between participants. As expected, the agreement rises when considering only one type of annotation (e.g. thumbs up only) but these results still do not indicate a high level of agreement among participants in terms of their system feedback.

Table 5.7 shows the Pearson product-moment correlation coefficient between a participant's different types of feedback for a given stream. There is a strong positive correlation between the four survey measures but none of these correlated with the document feedback.

In Table 5.8 we see the average ratings for all $< Topic, Algorithm >$ combinations at stream level. Each value is the average between the rating given by two users. The `control` algorithm, which randomly samples the labelled documents, was rated on average higher than all other sentiment filtering algorithms. Two algorithms, `negneu` and `posnegneu`, performed significantly worse than the control algorithm for the 16 topics (2-tailed, $p < 0.05$). All algorithms were rated above the midpoint in the scale (2.5), indicating that overall, the participants were satisfied with the system performance.

Examining Table 5.9, we can see that using document-level feedback only one algorithm was rated higher than the `control` algorithm, namely the `pos` algorithm. The `posneg` algorithm performs worst but surprisingly for no algorithm do we see a statisti-

Topic	pos	neg	neu	posneg	posneu	negneu	posnegneu	control	mean
48	5	3.5	5	4	4	3.5	3.5	4.5	4.125
15	2.5	3.5	4.5	3.5	4.5	3	5	4.5	3.875
13	3.5	3.5	3	3.5	4.5	3.5	3.5	5.5	3.8125
29	4	4.5	3.5	3.5	3.5	3.5	3.5	3	3.625
38	3.5	3.5	5	3	3	3	4	3.5	3.5625
11	4	4	3	3	3	3.5	4	3.5	3.5
31	4	2	3.5	3	3	4	3.5	4	3.375
17	3	3	4	2.5	3.5	3.5	3	2.5	3.125
1	2.5	2	3.5	3.5	4	2.5	2	3	2.875
9	3	3	2	3	3	3	2	2	2.625
39	2	2.5	3.5	2	3	1.5	2	4	2.5625
12	1.5	3	2.5	3.5	3	2.5	1	2.5	2.4375
4	1.5	1	4	2.5	3.5	2.5	2	2	2.375
43	2.5	2.5	2	2.5	2.5	2	1	4	2.375
6	2.5	4.5	0.5	2.5	1	1.5	2	3.5	2.25
2	3	3.5	1	2.5	1.5	0.5	1.5	2.5	2
mean	3	3.0938	3.1562	3	3.1562	2.7188*	2.7188*	3.4063	3.0313

Table 5.8: Overall algorithm-topic ratings (* denotes result significantly differently from control (2-tailed, $p < 0.05$))

cally significant different rating from the `control` (2-tailed, $p < 0.05$). With just two of the eight algorithms recording overall negative feedback scores, the participants' feedback at the document-level suggests they are as satisfied as they state in their surveys.

Focusing on the document-level, we wish to ascertain whether document-topic sentiment is an independent variable with respect to document feedback. To do this we use a measure for goodness-of-fit, Pearson's chi-square test. Chi-square compares the expected matrix of occurances between two (or more) variables to the observed counts to determine the probability of their independence. In our experiment, each item of document-level feedback corresponds to a sentiment label for that $< Document, Topic >$ pair. We consider three document types: *positive*, *negative* and *neutral*. Documents which are *mixed* are so few in number, and the label agreement so low, that it would be difficult to draw any conclusions, so we exclude them. There are also three possible document feedback values: *thumbs up*, *thumbs down* and *no annotation*. We thus have a 3-by-3 matrix of observed frequencies. See Figure 5.3 for the breakdown in document feedback per sentiment type.

We find a significant degree of dependence between feedback type and document-topic sentiment ($p < 0.05$). Examining the observed and expected counts, we can see where our observations deviate from the expected values (see Table 5.10). Overall, negative documents received more annotations than expected, while neutral documents received

Topic	pos	neg	neu	posneg	posneu	negneu	posnegneu	control	mean
31	0.5	0.0833	0.6389	-0.0556	0.0833	0	0.3333	0.25	0.2292
48	0.4167	0.0833	0.0417	0.0833	0.25	0.0833	0.2083	0.625	0.224
38	0.3056	0.3333	0.1667	0.0417	0.3333	0	0.2083	0.0417	0.1788
15	0.2083	0.0417	0.1250	0.2083	0.0833	0.5000	-0.0417	0.1667	0.1615
13	0.3333	0.0833	-0.125	0.1667	0.5556	-0.125	-0.1667	0.375	0.1372
11	0.1667	0.2083	0.0833	-0.25	0.3333	0.125	0.0417	0.25	0.1198
17	0.0417	0.125	0.2917	-0.125	-0.0833	0.125	0.25	-0.0417	0.0729
29	0.4167	0.2083	0.0556	-0.0417	-0.6667	0.3333	0.2222	-0.0833	0.0556
1	-0.0417	0.4583	0.125	-0.2083	-0.125	0.3333	-0.2917	0.125	0.0469
4	0.3333	0.0417	0.25	0.1667	-0.1667	0.1111	-0.6667	-0.2083	-0.0174
9	0.0417	-0.4167	0	0.3333	-0.0833	-0.0833	0.125	-0.0833	-0.0208
39	0.25	-0.1667	-0.4167	0.125	-0.1667	0.0417	-0.1111	0.2083	-0.0295
12	-0.25	0	-0.5	0.0417	-0.0833	0.25	-0.25	0.1667	-0.0781
43	0.125	-0.0417	-0.5	-0.4167	-0.1667	0.0833	0	-0.0417	-0.1198
6	-0.2083	-0.1667	-0.3333	-0.5833	0.25	-0.0833	0	0	-0.1406
2	-0.375	-0.4583	0.25	0.0417	-0.0833	-0.5	0.0417	-0.4167	-0.1875
mean	0.1415	0.026	0.0095	-0.0295	0.0165	0.0747	-0.0061	0.0833	0.0395

Table 5.9: Algorithm-topic ratings inferred from document-level feedback; no algorithm performs significantly different to the control algorithm (2-tailed, $p < 0.05$)

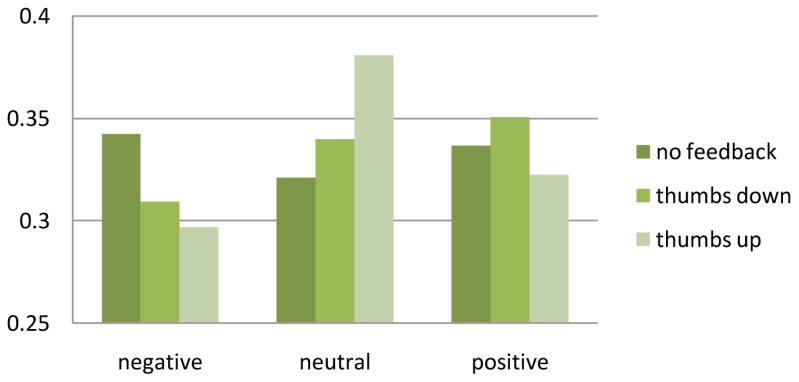


Figure 5.3: Feedback proportion for document sentiment type

Sentiment		no annotation	thumbs down	thumbs up	total
negative	Observed Count	435	255	290	980
	Expected Count	405.46	262.86	311.67	980
	% difference	+7.28	-2.99	-6.95	
neutral	Observed Count	408	280	372	1060
	Expected Count	438.56	284.32	337.12	1060
	% difference	-6.97	-1.52	+10.35	
positive	Observed Count	428	289	315	1032
	Expected Count	426.98	276.81	328.21	1032
	% difference	+0.24	+4.4	-4.03	
	Observed Count	1271	824	977	3072
	Expected Count	1271	824	977	3072

Table 5.10: Document sentiment and document-level feedback are associated according to chi-square ($p < 0.05$)

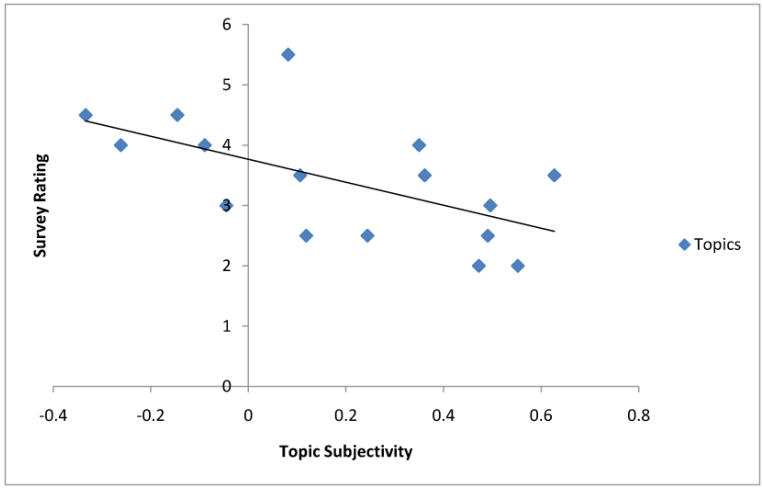
less feedback than expected. Neutral documents received disproportionately more thumbs up annotations, while positive documents received disproportionately more thumbs down.

We also performed chi-square tests for independence between feedback and algorithms, though neither the document-level feedback nor the survey feedback proved to be associated with respect to the algorithms ($p < 0.05$).

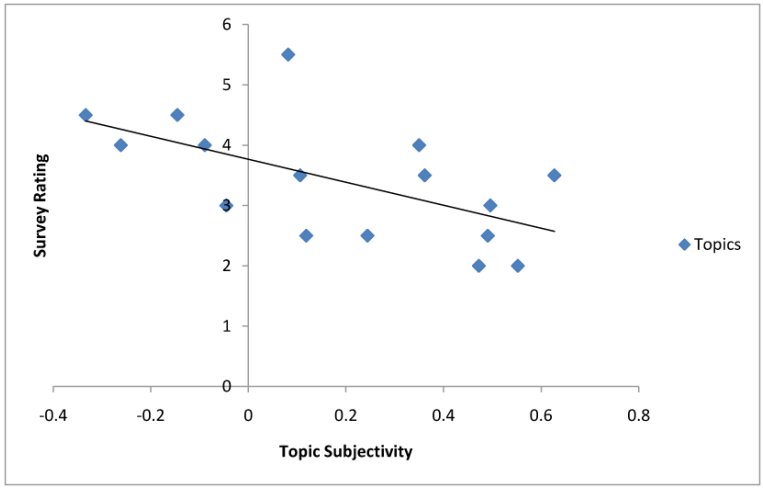
5.3.2 Topics

Recalling the measures of topic subjectivity and sentiment we defined in Section 5.2.1, we examine the relationship between these values and participant feedback. If we compare the overall rating for the `control` streams for each topic with topic subjectivity, we see a significantly negative correlation ($r = -0.59$, 2-tailed, $p < 0.05$). The higher the prevalence of subjective documents for a topic in the distribution, the lower participants rated the `control` stream for that topic (see Figure 5.4). We see the same correlation in our `control` streams between topic subjectivity and $NetFeedback$ (i.e. difference in proportion of positive and negative feedback) for those streams ($r = -0.55$, 2-tailed, $p < 0.05$). This correlation does not hold true for the *interesting* or *insightful* ratings. However, we observe an almost identical negative correlation for *informativeness* ($r = -0.57$, 2-tailed, $p < 0.05$).

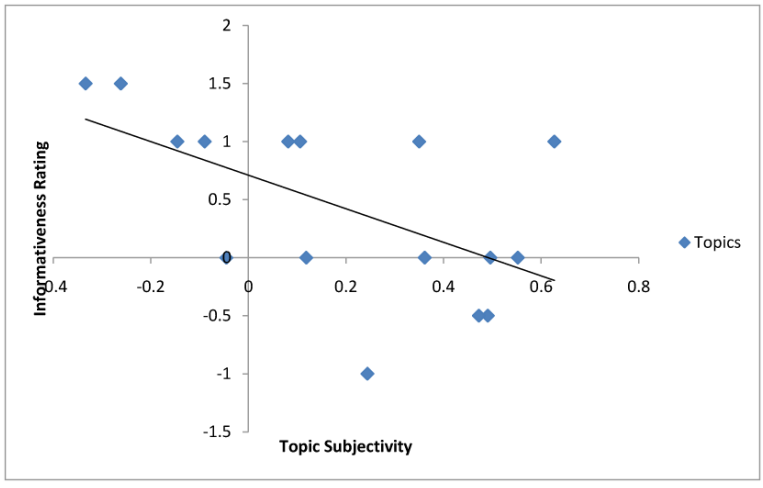
We found $NetFeedback$ score for the `pos` algorithm to be negatively correlated with the overall topic sentiment ($r = -0.62$, 2-tailed $p < 0.05$). The more positive topics were



(a) Overall Rating Vs Topic Subjectivity ($r = -0.59$, 2-tailed, $p < 0.05$)



(b) Document Feedback Vs Topic Subjectivity ($r = -0.55$, 2-tailed, $p < 0.05$)



(c) Informativeness rating Vs Topic Subjectivity ($r = -0.57$, 2-tailed, $p < 0.05$)

Figure 5.4: For the `control` stream, overall rating and document feedback was negatively correlated with topic subjectivity. This was likely due to perception of *informativeness*.

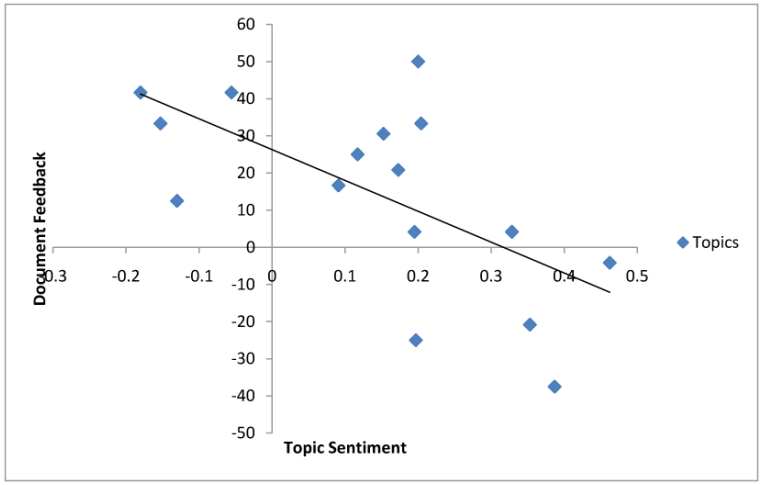


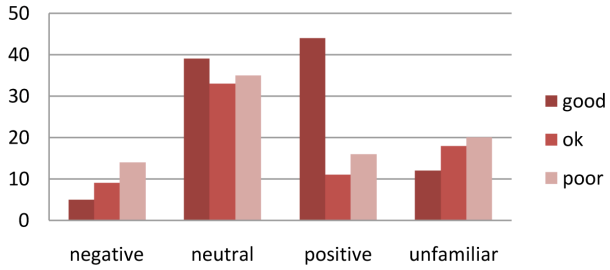
Figure 5.5: Feedback for the `pos` algorithm becomes more negative for topics with more positive sentiment ($r = -0.62$, 2-tailed, $p < 0.05$)

in our labelled corpus, the more negatively participants reacted to the `pos` streams (see Figure 5.5). In this and the preceding document feedback correlation, these significant correlations still hold true if we only consider thumbs up annotations, yet not if we just consider thumbs down annotations.

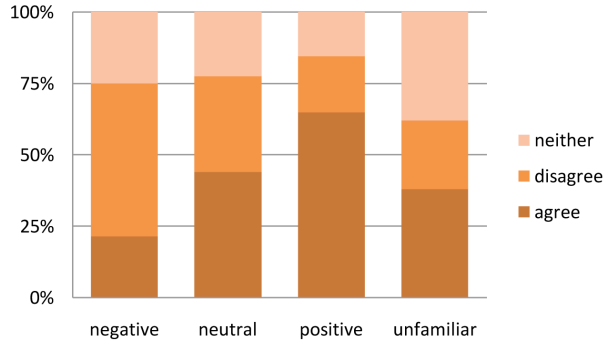
5.3.3 Prior Sentiment

We asked participants before each topic what their own personal opinion, or *prior sentiment*, was towards that topic. Participants answered that they had either primarily *positive* opinions towards the topic, had primarily *negative* opinions about the topic, were familiar with the topic but had neither positive nor negative opinions about the topic (*neutral*), or that they were *unfamiliar* with the topic. Examining the overall breakdown in document-level feedback, immediately obvious is how seldom participants declared themselves as negative for topics, representing just 10.94% of all streams, see Figure 5.6a.

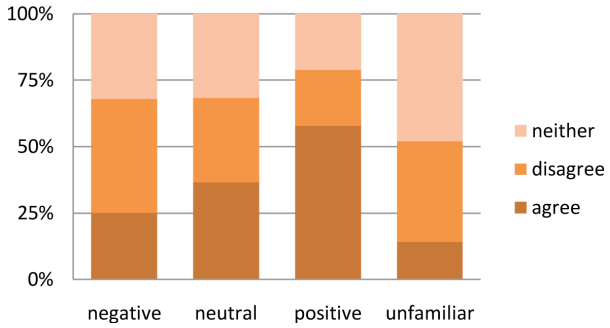
We find that each of the survey feedback questions were dependent on participant prior sentiment ($p < 0.001$) (see Figure 5.6). The most prominent pattern is the positive overall rating that participants gave to streams where they were positive about the topics. This appears to be consistent across the ratings, with roughly 60% of positive participants agreeing that the streams were informative, interesting and insightful. The opposite is true for those negatively predisposed, with approximately 50% disagreeing with these



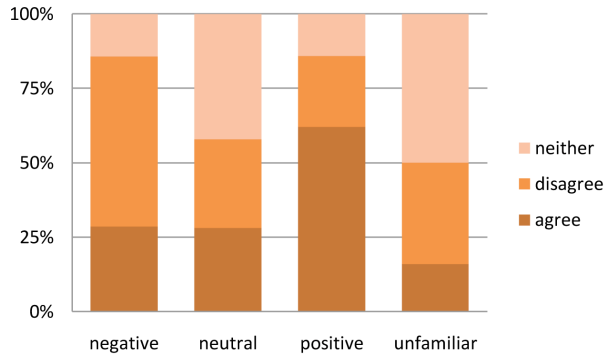
(a) Overall



(b) Informative



(c) Insightful



(d) Interesting

Figure 5.6: $< Participant, Topic >$ prior sentiment was found to be linked to each of our survey measures ($p < 0.05$)

Participant Sentiment	Algorithm							
	control	neg	negneu	neu	pos	posneg	posnegneu	posneu
Negative	-0.04	0.22	0.08	n/a	-0.19	-0.17	-0.35	-0.21
Neutral	0.08	0.03	-0.04	-0.10	0.06	0.05	0.03	0.08
Positive	0.37	0.02	0.25	0.38	0.12	0.14	0.11	0.10
Unfamiliar	-0.08	-0.07	-0.04	-0.03	0.07	0.01	-0.11	-0.02

Table 5.11: Mean $NetFeedback$ scores for algorithms and prior participant sentiment

Participant Sentiment	Algorithm							
	control	neg	negneu	neu	pos	posneg	posnegneu	posneu
Negative	3.00	4.33	1.75	n/a	1.33	2.67	2.50	2.75
Neutral	3.43	2.82	2.91	2.77	3.17	3.00	2.67	2.93
Positive	3.70	3.00	3.50	4.09	3.56	3.75	3.43	3.89
Unfamiliar	3.00	3.43	3.00	2.83	2.75	2.14	2.43	2.75

Table 5.12: Mean overall ratings for algorithms and prior participant sentiment

three descriptions.

In Table 5.11 and Table 5.12, we can see the mean feedback broken down by participant prior sentiment and algorithm for document-level feedback and overall rating respectively. Although the sample sizes are small, some patterns are apparent. For one, positive participants provided the most positive feedback for both measures. The feedback for the unfamiliar and neutral participants is more moderate. The feedback for the negative participants is most intriguing, with a wide variance across algorithms. In particular, negative participants rated the negative stream the highest and the four streams containing positive sentiment (`pos`, `posneg`, `posneu`, `posnegneu`) much lower. This pattern is not as evident for the positive participants, though we do see that the positive participants rated the negative stream lowest in terms of document-level feedback and overall rating.

To futher investigate the prior sentiment feedback patterns, we can look at the log odds ratios for thumbs up and thumbs down feedback, given the various participant prior sentiment and document sentiment combinations (see Table 5.13). There are three significant patterns we observe. First, the positive participants were twice as likely to annotate a document *thumbs up* as other participants, and this was consistent across document types. Conversely, participants were twice as likely as others to *thumbs down* a document when they were negative towards a topic, *except* when the document is negative. Lastly, par-

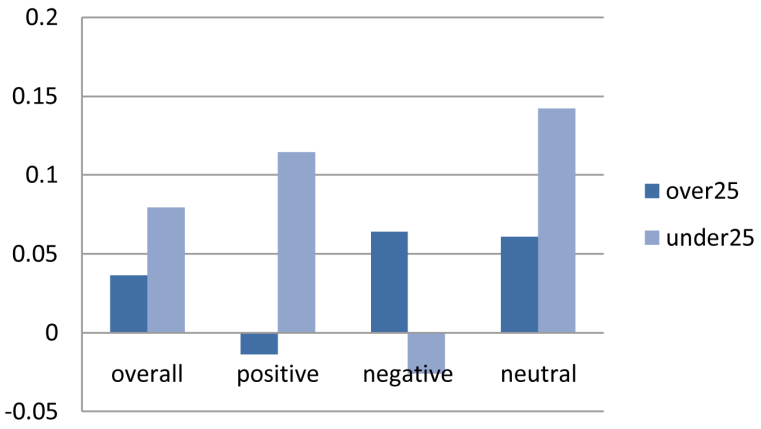
Participant Sentiment	Document Sentiment	Log Odds	
		Thumbs Up	Thumbs Down
Positive	all	0.3**	-0.12*
	positive	0.22**	-0.03
	negative	0.35**	-0.14
	neutral	0.33**	-0.2*
Negative	all	-0.02	0.29**
	positive	-0.1	0.35**
	negative	-0.12	0.09
	neutral	-0.03	0.4**
Neutral	all	-0.02	0.06
	positive	0.14*	-0.04
	negative	-0.06	0.17*
	neutral	-0.13*	0.05
Unfamiliar	all	-0.35**	-0.15*
	positive	-0.52**	-0.18*
	negative	-0.37**	-0.16
	neutral	-0.21*	-0.11

Table 5.13: Log odds ratios for document-level feedback type with respect to user prior sentiment and document sentiment. Significance according to chi-square at $p < 0.05$ (*) and $p < 0.001$ (**)

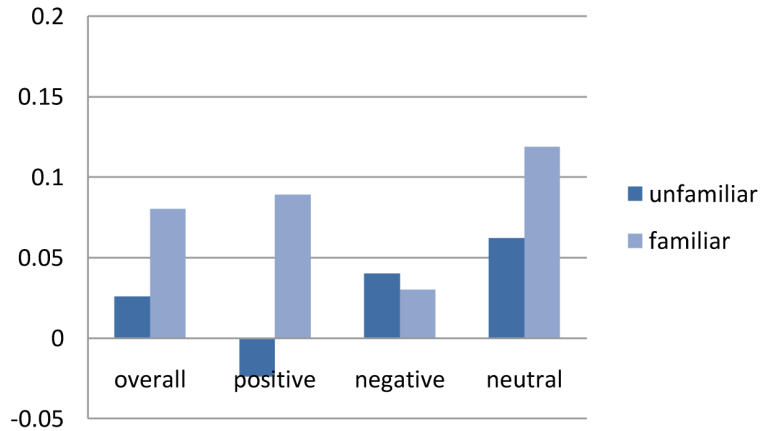
ticipants who were unfamiliar with topics were less than half as likely to provide thumbs up feedback. This pattern is particularly true for positive documents; participants who were unfamiliar with the topic were three times less likely to thumbs up a positive document. For neutral participants, there are fewer significant patterns, and the effect sizes are smaller.

5.3.4 Participant Profiling

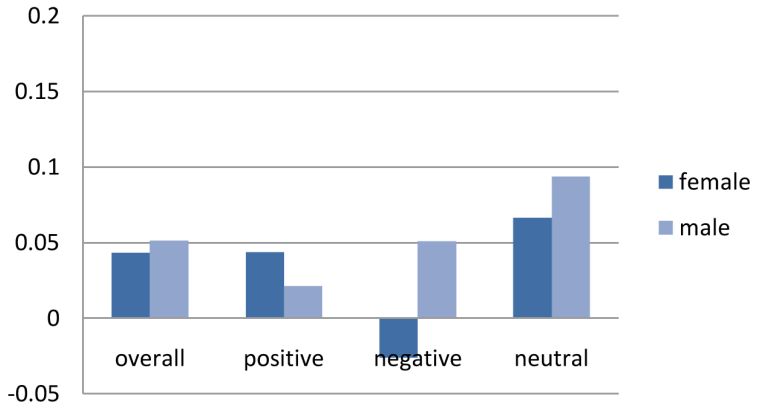
Lastly, we find significant associations between three aspects of the user profiles and their document-level feedback (2-sided, $p < 0.001$) (see Figure 5.7). For age, we see younger participants were generally more positive in their feedback, apart from negative documents which older participants preferred. Those familiar with Twitter search were more positive towards neutral and negative documents. The primary difference in gender is between positive and negative documents, with females preferring positive documents, and males preferring negative.



(a) Age



(b) Twitter Search Familiarity



(c) Gender

Figure 5.7: We observed associations between age, gender and task familiarity and document-level feedback (2-sided, $p < 0.001$; here illustrated as $NetFeedback$ scores)

5.4 Discussion

In this section, we discuss our experimental observations with reference to our research questions. We also consider our methodology and the relative strengths of our evaluation measures.

5.4.1 Feedback Mechanisms

Throughout our evaluation we use two fundamental methods of evaluation: surveys and explicit document-level feedback. How accurately are these mechanisms capturing the feedback required for us to assess our research questions? We asked our participants to give feedback based on what they saw as desirable in the real-time information access system. Without tying the participants to any objective criteria, it is perhaps not surprising that the agreement between annotators is so low. Participants showed almost no agreement in terms of their overall rating for a given $< Topic, Algorithm >$ stream. In fact, α is so low as to suggest that there is no agreement among participants beyond what is expected by chance. We consider three possible reasons for this:

1. The participants were rating the streams based on different criteria.
2. There were too few documents in the stream and participants' views were skewed by reactions to specific documents.
3. Participants were strongly influenced by external factors.

For reason (1) above to be solely responsible for lack of agreement, we would have to assume that the criteria being used by the annotators were so different as to cause as much disagreement as agreement. We think this unlikely, particularly as there was some (albeit modest) level of agreement for document-level feedback. The higher level of document-level agreement also indicates that reason (2) is a likely cause. Two streams that were assigned the same algorithm and topic will likely still contain few common documents. With just 12 documents in the stream, participants may have characterised the stream based on these documents rather than building up an overall impression of the nature of the content in the stream. The patterns we observe throughout our evaluation with respect

to stream survey ratings do suggest that the mean ratings are indeed measuring an overall impression of the stream, despite the disagreement between the individual participants. There is also significant indication to suggest that (3) above impacts participants' ratings. See Section 5.4.4 for a further discussion on participant profiling.

5.4.2 Topics

The strong positive corelation between sentiment and subjectivity amongst the labelled documents for our topics is intriguing. With a sample size of 50 topics, it is reasonable to assume that this pattern is not unique to our corpus. So the question remains — why for trending topics, when there is a higher degree of subjective content, is that content more likely to be positive? The decomposition of the topics into their categories is revealing.

Only one of the *Companies* and two of the *Politics and Government* topics had a positive sentiment score. For these topics, a higher degree of subjectivity indicated a higher degree of negative sentiment. These topics were typically trending due to topical controversy, and it appears that Twitter users were expressing their criticism of them. Despite this pattern, most other topics evoked a greater degree of positivity when the content was more subjective. This was particularly the case where users declared their support for topics, for example sports teams, musical acts and products. There appears to be a division between topics related to current affairs, which attract critical content, and other areas, which evoke a more positive response. This has important ramifications for applications that attempt to measure a real world absolute sentiment value for a given topic. A model for correcting for topic-category skew could help normalise the Twitter-based sentiment scores and allow them to be understood and used alongside other measures; in these applications identifying the true sentiment in a population beyond Twitter is often the challenge.

One pattern we found interesting was the negative correlation between topic subjectivity and overall stream rating. The streams for these high-subjectivity topics were rated poorly by our participants and, as we saw in Figure 5.4, this is likely caused by a perceived lack of informativeness for these topics. The *Entertainment* topics were amongst the most subjective. We speculate that these topics tend to be more frivolous and less substantive,

perhaps lacking appeal outside a niche interest. Nonetheless, these correlations do not hold true for how interesting or insightful streams are — is perceived informativeness a deal-breaker for real-time users?

5.4.3 Algorithms

None of our filtering algorithms performed significantly better than the `control` algorithm and, only in two cases did they perform significantly worse. Overall, this part of the study was inconclusive and does not support our hypothesis that sentiment filtering algorithms perform significantly differently from the control. There is sufficient evidence to suggest that with a more extensive evaluation a pattern may be observed, reducing the risk of experiencing a type II error. For example, in the course of our evaluation we noticed that other pairwise comparisons between algorithms yielded significant differences. As we noted in Section 5.4.1, we suspect that longer streams may yield more revealing results with respect to stream-level evaluation.

Regarding topic qualities, as we saw, our `pos` algorithm does poorly for positively regarded topics, and our `control` algorithm does poorly for subjective topics. We note that these suggest certain sentiment-topic interactions, but without further study it is difficult to draw any conclusions.

5.4.4 Prior Sentiment

When dealing with sentiment, an inherently subjective concept, it is important to consider the effect that a person's personal state and world view has on their perception. The fact that we have linked participant prior sentiment to each of the survey measures demonstrates the significant role played by the views of the user. From the outset, the obvious question to ask then is, do people prefer documents containing sentiment which align with their own view? The stream feedback suggests that this may be the case, at least for positive and negative $< Participant, Topic >$ combinations.

However, looking at net document-level feedback, the results in Figure 5.8 are surprising, and show our suspicions about alignment between document sentiment and personal sentiment to perhaps be ill-founded. In fact, positively and negatively predisposed par-

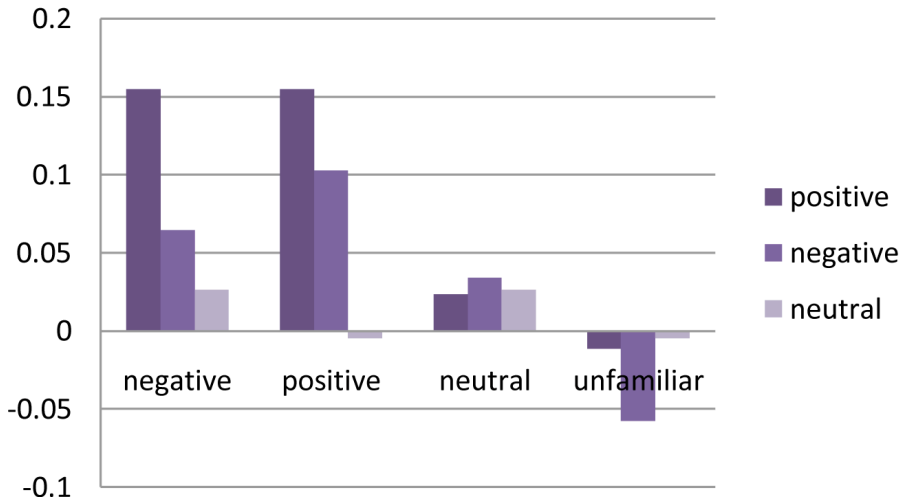


Figure 5.8: *NetFeedback* score for document sentiment type, grouped by participant prior topic sentiment

ticipants behaved similarly in terms of *NetFeedback*, but differently to the neutral and unfamiliar participants. Those who held an opinion were more likely to be positive about both positive *and* negative documents, than those who had described themselves as neutral or unfamiliar.

When we consider the thumbs up and thumbs down annotations in isolation however, we see a different pattern. It is clear that each prior sentiment category exhibits a distinct pattern of feedback, reinforcing what we had previously seen in survey measures. However, in some cases we see that by isolating feedback for document sentiment types, we can reveal some patterns that demonstrate that feedback is dependent not just on the participant's prior sentiment, but also on the sentiment of the document itself. The best example of this is the feedback from negative participants which clearly discriminates between documents on the basis of document sentiment. Interesting also, is the characterisation of unfamiliar participants; they were considerably less likely to offer positive feedback to the system. Perhaps this is due to the lack of contextual knowledge with which they approach the task, leading them to feel unqualified to offer feedback.

There is clearly a complex system of interdependencies here. It appears that participant prior sentiment has a significant impact on participant feedback, and that this is consistent across all measures. We note also that in some cases, these effects are not consistent between document sentiment types and vary significantly. In particular, using document-level sentiment, we can see how, although positive and negative participants

appear to behave inversely (positive participants more likely to thumbs up, negative participants more likely to thumbs down), the manifestation of these behaviours with respect to document sentiment is not necessarily similar.

5.4.5 Participant Profiling

Similar to participant prior sentiment, in looking at other aspects of participants' profiles we observe how age, task familiarity and gender affect sentiment perception. Each of these three profile attributes were observed to be associated with document-sentiment feedback. Our under 25s liked the positive documents; 25s and older disliked the positive documents, instead liking negative documents. Similarly those familiar with microblog search preferred positive documents to negative documents; those who were unfamiliar preferred negative documents and disliked positive documents. The difference was less pronounced for gender, though our female participants rated negative documents particularly low. Collectively these results, alongside the predisposition results, make a compelling argument for exploring the idea of building real-time information systems with a sentiment-based recommender component which learns a user-sentiment profile and presents documents accordingly.

Our sample size, while significant, is still relatively small so making general demographic conclusions is outside the scope of this study. For example, our small group of five under 25s may be subject to common external variables such as job role or technical expertise, which could cause an observed correlation in behaviour, unrelated to their age.

5.5 Conclusion

In this chapter we have explored simulated real-time scenarios with 16 users, for 16 diverse topics. We have devised and evaluated an experiment that used user feedback to determine the effect of sentiment with respect to users, topics, algorithms and the documents themselves. We used both explicit system feedback, and survey feedback, alongside topic measures and user profiles to perform our analysis. We used Pearson correlations to identify significant relationsips between numerical data and, where the data is categorical, we used chi-square to assess variable independence.

An analysis of our topics and labelled data revealed interesting patterns, including a positive relationship betwen topic subjectivity and topic sentiment. We identified a significant association between document-level feedback and document sentiment. However, our sentiment filtering algorithms in most cases did not demonstrate any conclusive deviation in feedback from a `control` algorithm. We found significant differences in feedback with respect to participant profile and prior sentiment across all feedback measures. We also found that, in some cases, these patterns were related to the sentiment of the documents themselves. We conclude that user profile and prior sentiment has a significant association with their feedback and perceived content quality. We conclude also that a significant role is played by document-level sentiment with respect to participant sentiment, although this pattern is more complex than simply participants preferring documetns which were aligned with their view.

The results we observe here are encouraging in that they support the hypothesis that sentiment plays a vital role in real-time information access. However, as mentioned in the introduction to this chapter, evaluating with simulated scenarios has its limitations. In the next chapter, we put participants in live real-time scenarios and perform a deep exploration of two topics, one in the area of *Entertainment* and the other in *Politics and Government*.

Chapter 6

Real-time User Studies

6.1 Introduction

In this chapter we reach the culmination of our research. Having evaluated supervised learning techniques for microblog sentiment analysis, and having made observations during simulated real-time microblog search scenarios, we now deploy the Channel S system in a series of real-time user studies. Using real world topics with real-time data and user feedback we evaluate our sentiment-focused hypotheses.

At this stage it is useful to recall our research questions:

- *Do sentiment-based algorithms differ significantly from a baseline sampling approach?*
- *Do users' demographics and preferences significantly effect their perception of sentiment? Which types of sentiment have the most profound impact?*
- *Is sentiment a predictor of whether individual documents will be regardly as important by users?*

As we saw in Chapter 2, these questions lead us to focus on examining the effect of sentiment on the search task at three levels: the sentiment of (i) the stream, (ii) the document and (iii) the user. Throughout our experiments we examine these factors as independent variables. Then, from user feedback, our dependent variable, we perform analysis which allows us to address the above research questions.

It is non-trivial to identify real-time scenarios which are suitable for our evaluation. As a medium-sized user trial, significant resources and organisation are required to run the studies. We therefore need to commit ourself to a small number of real-time topics and be able to prepare in advance. Perhaps some of the more interesting real-time topics concern breaking news stories. These spontaneous topics are however inherently unpredictable, and organising real-time laboratory user trials around these would be difficult, if not impossible.

Fortunately, there are other types of real-time topic what we can use for our studies: scheduled events. Scheduled events may be for example concerts, sports matches, debates, television programmes, presentations or conferences. Unlike spontaneous topics, scheduled events typically have a predictable structure, and defined beginning and end points. This allows us to make assumptions about the timeline of the event. We can also make assumptions about the nature of relevant microblog content through observing previous occurrences of similar events.

In the experiments in this chapter we choose two real-time topics: the X Factor, 2010, and the Leaders' Debate during the Irish General Election, 2011. The X Factor is a popular television programme in terms of social media, attracting many thousands of microblog posts per episode. The content is also emotionally charged, containing reaction to contestant performances and expressions of support or derision for the contestants and the judges in the show. The Leaders' Debate, although more serious in nature, similarly attracts statements of support and derision from microblog communities, reacting to topical issues and individual debate performances. The two topics are similar enough in structure and nature to allow us to replicate the same experiment on each, yet different enough that we can contrast and compare the role of sentiment in one, an entertainment event, and in the other, a political event.

In the next section we detail our experimental methodology including our laboratory set-up, ethical considerations and configuration of the Channel S system. In Section 6.3 we present our results and in Section 6.4 we discuss these results with reference to our research questions. In Section 6.5 we digress from our evaluation to present the GE11 Twitter Tracker system which we developed for political sentiment monitoring during the General Election. Finally, we conclude the chapter in Section 6.6.

6.2 Methodology

Our methods follow standard methodology for a formal *repeated measures* user study. We conducted our experiment three times, replicating the same experimental circumstances in each. In this section, we discuss our experimental set-up, ethical considerations, the configuration of our experimental system, and data and measures for evaluation.

6.2.1 Experimental Set-up

The contraints of our research dictate that the data we gather is from real-time user behaviour as they use the search system. We therefore set up our experiment in a shared space, where a number of users can use the system concurrently while observing the topic event live on a large shared screen. We determined that approximately 20 users is appropriate for this type of experiment, allowing us to capture sufficient data yet still remaining manageable in a shared environment.

The X Factor is a singing and performance contest on ITV television. Each week, the contestants perform and through phone voting and a panel of judges, a contestant is eliminated. The two shows we use for data capture, took place on Saturday the 11th of December and Sunday the 12th of December. These were the penultimate and final shows of the series respectively. Cognizant that a significant portion of the viewership of the X Factor is adolescent and younger, our first run of the experiment is with a younger group of 17 participants, chiefly aged from 18 to 20. On the second day, our group was comprised of 18 research staff and students.

The Leaders' Debate is a televised debate whch took place between the leaders of the five primary political parties in Ireland. It was a focal point of the Irish General Election campaign period and the only time the five leaders participated in such a debate. At the time, Ireland's government had dissolved prematurely amid public dissatisfaction with its handling of the economic crisis. There was therefore much focus on the impending election as a change in government was anticipated. The Leaders' Debate took place on the 14th of February, and the experiment was run with a diverse set of 21 participants consisting of university research staff, members of the student body, and their friends. See Figure 6.1 for photographs of the studies in progress.

(a) The X Factor, 11th December, 2010

(b) The X Factor, 12th December, 2010

(c) The Leaders' Debate, 14th February, 2011

Figure 6.1: User studies in progress

For each experiment, participants were required to arrive 30 minutes in advance of the event beginning. We provided each participant with instructions for logging into Channel S, a set of written instructions, relevant ethics information and a booklet of surveys to be completed during the experiment. We also gave them a survey capturing demographic information and familiarity and opinion towards the topic. The surveys used were the same as the surveys used in our simulated experiment (Chapter 5) with some minor changes.

Once all participants had arrived, we gave the participants a short presentation on how to use the system and provide feedback. We also provided information about microblogging as well as the background to the topic for the benefit of those who were not familiar. Participants were then allowed to use the system until the start of the event, though only feedback during the event was used in our evaluation. By allowing particpants to familiarise themselves in such a way, we minimize any learning effect.

We assigned each user an identification number. This number randomly put each participant in one of four groups. When the event began, each participant was allocated a sentiment algorithm corresponding to their group by the system. At intervals of 15 minutes, the system prompted participants to complete a survey based on the stream for the preceding 15 minutes. The algorithms were then rotated and each group was assigned a new algorithm. Each group experienced each algorithm an equal amount of times and in the same order. Teh aglorithms used were subset of those we examined in our simulated real-time experiment in Chapter 5.

At the end of the experiment users completed a closing survey and forms and surveys were collected. See Appendix C for all documentation related to the experiment.

6.2.2 Ethics

In preparation for these studies, we compiled a submission to the university Research Ethics Committee notifying them of our intention to run a user study. As our experiments were not invasive, required only a minimal amount of personal information, and posed no risk to the participants, our experiments were classified as elligible for low risk ethics approval. There were a number of components to our ethics submission:

- *Notification Form*: In this form, we describe our methodology and research goals. We also describe how participants are recruited, how their anonymity and confidentiality is respected and confirm that they are not exposed to any risk.

- *Plain Language Statement*: This is the first material that the participants read before the experiment commences. The plain language statement explains what is expected of the participants in plain, non-technical terms. It also assures them that their data is stored anonymously, and that the experiment poses no risk to them.

- *Informed Consent Form*: In completing the informed consent form, participants acknowledge that they have read the plain language statement and that they understand what is required of them. They also acknowledge that they are participating voluntarily and are free to leave the study at any time.

- *Participant Questionnaires*: Lastly, our ethics submission required that we submit copies of all of our participant surveys to ensure that they are consistent with our study, as we described.

Our ethics submission was approved by the university prior to running our experiments. The relevant ethics materials are contained in Appendix C.

6.2.3 Evaluation Measures

There are a number of different aspects to our evaluation, so our measures must be chosen appropriately. The three types of data we use in our evaluation are: (i) document-level feedback, (ii) stream-level feedback from periodical surveys, and (iii) participant profile data from the introductory surveys. When document-level feedback is aggregated over a set of documents, as in the previous chapter, we refer to the mean $NetFeedback$ score for a set of documents, D:

$$NetFeedback(D) = \frac{|D_{thumbsup}| - |D_{thumbsdown}|}{|D|} \tag{6.1}$$

For stream-level feedback, our primary data is a 7-point numerical scale where participants rate the stream from *poor* (1) to *excellent* (7). Secondly, we use 5-point Likert agreeability scales for *interestingness*, for *insightfulness* and for *informativeness*. In order to simplify

our evaluation, we conflate these 5 point scales to 3-points scales (simply *agree*, *disagree*, *neither*). Similarly we conflate the overall scale to *poor* (1,2,3), *ok* (4) and *good* (5,6,7) where categorical rather than numerical data is required.

Initially we examine the effect of varying the sentiment in the streams. Our experiment is a repeated measures experiment with four experimental conditions, four stream filtering algorithms: (i) positive documents only (`pos`), (ii) negative documents only (`neg`), (iii) positive and negative documents only (`posneg`) and (iv) random sampling (`control`). In evaluating, we require a test that tells us whether the distribution in participant feedback varies significantly under our four experimental conditions. We use the general linear model for repeated measures to compare the feedback distribution under the four conditions, from survey data, and aggregate document-level feedback. This allows us to compare the feedback for the four conditions to see if altering the filtering algorithm is inducing a significant difference in feedback distribution.

In addition, as one of our conditions is a baseline control condition, we compare the feedback from each of our other conditions to the control condition using a t-test. We have not predicted a direction for our distributional shift in our hypothesis; we are testing if the different experimental conditions produce a shift in either direction. The same subjects were used to produce the values for each survey and each data point in each set has a corresponding point in the comparison distribution. Out t-test is therefore 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, which corresponds to attribute differences between participants. We can thus test whether user prior opinion or demographic details have a significant impact on the main effect. See Table 6.1 for the sample sizes for various participant attributes. In dividing up our participants into different demographic groupings we have endeavoured as much as possible to use groupings which divide the participants as equally as possible.

At document level, each item of feedback has a number of categorical variables; each feedback action is associated with a participant and their profile attributes, the sentiment of the document is identified by Channel S, and a feedback action (one of *thumbs up*, *thumbs down* or *no annotation*). We perform a series of tests for independence between

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
Education	bachelor's degree or higher	17	18
	no degree	17	3
Prior Sentiment	positive	16	4
	negative	9	4
	neutral	7	9
	unfamiliar	3	4

Table 6.1: Participant sample sizes for profile attributes

these categorical variables using Pearson's chi-square test. Throughout we use odds ratios to describe the effect size in binary categorical associations. Statistical significance is reported at $p < 0.05$ (*) and $p < 0.001$ (**) where relevant.

6.2.4 Sentiment Analysis Configuration

In these experiments, we consider real-time microblog search more deeply that we have in our earlier, simulated real-time study. Previously we had considered the sentiment topic and query topic to be one and the same. This simplifying idea perhaps ignores a significant volume of sentiment which relates to entities related to the topic, but not necessarily the same as the topic itself. In preparation for these experiments, exploring related content confirmed this. For example, during the X Factor, many authors tagged their posts "#xfactor", declaring explicitly that the content was relevant to the topic. However much of the content was centred around discussing entities at a sub-topic level. Authors discuss songs, performances, contestants and judges, and rarely explicitly refer to the show as a whole. If we were to take our previous assumption, it would be difficult to consistently interpret how the sentiment in this content relates to the X Factor in general. Similarly in the election, people are more likely to discuss policies, parties and candidates rather than talk about the election, or indeed the debate as a whole.

For these reasons, the sentiment targets we use for the X Factor are the judges and contestants. Similarly, for the election our sentiment targets are the party leaders and

	XF		GE11	
	Docs	%	Docs	%
positive	2,131	30.84	884	12.23
negative	3,640	52.68	2,716	37.58
neutral	843	12.2	3,628	50.2
mixed	296	4.28	153	2.12
Total	6,910		7,381	

Table 6.2: Labelled training documents for sentiment

their parties. During the training data creation phase, our annotators label documents with respect to these sentiment targets. 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 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.

The guidelines for labelling data used were similar to those used in our earlier experiments (see Appendix B). We take care to ensure sufficient diversity in our training examples. If we take our training data from a single point in time, there is a risk that that sample could display a particularly skewed sentiment distribution or a vocabularly specifically relevant to events at that time. To limit this effect, in the X Factor we take our sample data from two different shows earlier in the competition. For the Leaders' Debate we use labelled data from two separate weeks during the election campaign.

The breakdown of labelled documents can be seen in Table 6.2. We follow the same annotation methodology as in Chapter 4 and discard conflicting and duplicate annotations. Interestingly, we used two annotators who were familiar with sentiment analysis for labeling X Factor documents, resulting in an agreement of 0.78 (Krippendorff's α) for 3 classes: positive, negative and neutral. For the election, we used a group of annotators who were new to sentiment analysis, resulting in a labelled set of documents with an agreement of 0.48. This indicates a moderately high level of agreement for newly trained annotators, but a very high level of agreement for more familiar annotators. It is also possible that the sentiment expressed in X Factor data is more overt, and therefore easier to interpret consistently.

The `control` algorithm does not require any sentiment analysis as a random sample

of relevant documents is included. We consider relevant documents to be those which mention both the topic Twitter hashtag (#xfactor or #ge11), as well as a sentiment target. Our three sentiment-based algorithms (`posneg`, `pos`, `neg`) require that we can identify positive and negative documents to filter the stream of relevant documents. For the X Factor we tackled this using two separate binary classifiers. The first is trained to distinguish positive documents from negative, neutral and mixed sentiment documents. The second distinguishes negative documents from positive, neutral and mixed sentiment documents. Naturally, the positive classifier is used to identify documents for the `pos` algorithm and the negative classifier is used to identify documents for the `neg` algorithm. For the `posneg` algorithm, when each document is queued it is chosen from either the positive documents or the negative documents with equal probability. Documents which were classified as positive, *and* as negative, are marked unclear and discarded.

After the X Factor experiment had been completed, we determined that the altogether simpler architecture of a three-way classifier would attain the same performance. Indeed, for the Leaders' Debate we used a three-way classifer (positive, negative, neutral) and assigned documents to the algorithms in the same manner. In all cases, we aimed to maintain a rate of one document every 10 seconds for presentation to the user, which we have identified as an appropriate speed. On the rare occasions that there were insufficient documents of a particular type to satisfy the queue for a given stream, we used documents selected at random. This may mean, for example, that a negative or neutral document may be queued for the `pos` stream if insufficient positive documents are available. We deemed UI consistency from the participants' perspective to be important to the experiment across all experimental conditions.

Our feature vector consists of unigrams that occur in two or more documents in the training set. In Chapter 4, we saw the importance of the discriminability of sociolinguistic features. The tokenizer we use for our real-time trials is optimised for user-generated content so all sociolinguistic features such as emoticons (":-)") and unconventional punctuation ("!!!!") are preserved (Laboreiro et al., 2010). As before, we remove all topic terms, usernames and URLs to prevent any bias being learned towards these.

In Table 6.3, we see a moderately high accuracy for both our negative and positive classifiers for the X Factor. As before we train our SVMs with a linear kernel and cost

	negative	positive
Trivial	52.68	69.16
SVM	73.50	82.47
MNB	73.04	75.57

Table 6.3: Binary sentiment classification accuracies for X Factor data using 10 fold cross validation (Each classifier is trained to classify the target class from its complement e.g. {positive} vs {negative,neutral,mixed})

parameter set to 1. The Trivial classifier simply assigns all test instances the majority training label. As the SVM outperforms MNB in terms of accuracy, we use SVM in our live system. In Chapter 4 we saw MNB outperform SVM for microblog data. It is difficult to say why we do not observe the same effect here. We speculate it is due to the fact the topic focus is more narrow or that more training data available.

For GE11 data, we encountered a problem in employing our three class classifier. Due to the prevailing negative sentiment, our labelled data for the election contains comparatively few positive examples, just 12%. Neither an SVM nor a MNB classifier achieved an acceptable true positive rate for the positive examples. Using either of these classifiers, we would not be able to effectively identify positive examples for our pos and posneg algorithms as the learner biases towards the majority classes. To mitigate this effect, we evaluated a boosting approach that through iterative learning, upweights training examples from minority classes, thus improving recall for these classes. We used Freund and Schapire's Adaboost M1 method with 10 training iterations as implemented in the Weka toolkit[1] (Freund and Schapire, 1996). Following from this, we use an Adaboost MNB classifier which achieves 65.09% classification accuracy in 10-fold cross validation for three classes (see Table 6.4 for performance measures, and Table 6.5 for confusion matrices).

6.3 Results

In this section, we detail our experimental results. The results for the X Factor (*XF*) cover the two shows during which we collected data, and the data for the General Election (*GE11*) is from the Leaders' Debate. To allow us to contrast and compare the two topics,

[1]http://www.cs.waikato.ac.nz/ml/weka/

Classifier	Accuracy	True Positive Rate			F-score
		positive	negative	neutral	
Trivial	50.19	0	0	1	0.335
MNB	62.94	0.584	0.007	0.561	0.832
ADA-MNB	65.09	0.645	0.334	0.689	0.7
SVM	64.82	0.631	0.201	0.634	0.768
ADA-SVM	64.28	0.638	0.362	0.623	0.726

Table 6.4: Classification accuracies, per-class true positive rate and F-score for 3-way sentiment classification on GE11 data using 10 fold cross validation

Classifier	Document Label	Classified as		
		negative	neutral	positive
Trivial	negative	0	2,716	0
	neutral	0	3,628	0
	positive	0	884	0
MNB	negative	1,523	1,193	0
	neutral	606	3,020	2
	positive	217	661	6
ADA-MNB	negative	1,872	785	59
	neutral	936	2,538	154
	positive	243	346	295
SVM	negative	1,722	977	17
	neutral	793	2,785	50
	positive	243	463	178
ADA-SVM	negative	1,692	918	106
	neutral	807	2,634	187
	positive	212	352	320

Table 6.5: Confusion matrices for three-way sentiment classification on GE11 data

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

Table 6.6: Mean feedback for sentiment filtering algorithms (Significant differences noted for each measure according to within-subjects test for main effect using the general linear model; significance for individual distributions determined with respect to the corresponding control distribution (paired, two-tailed))

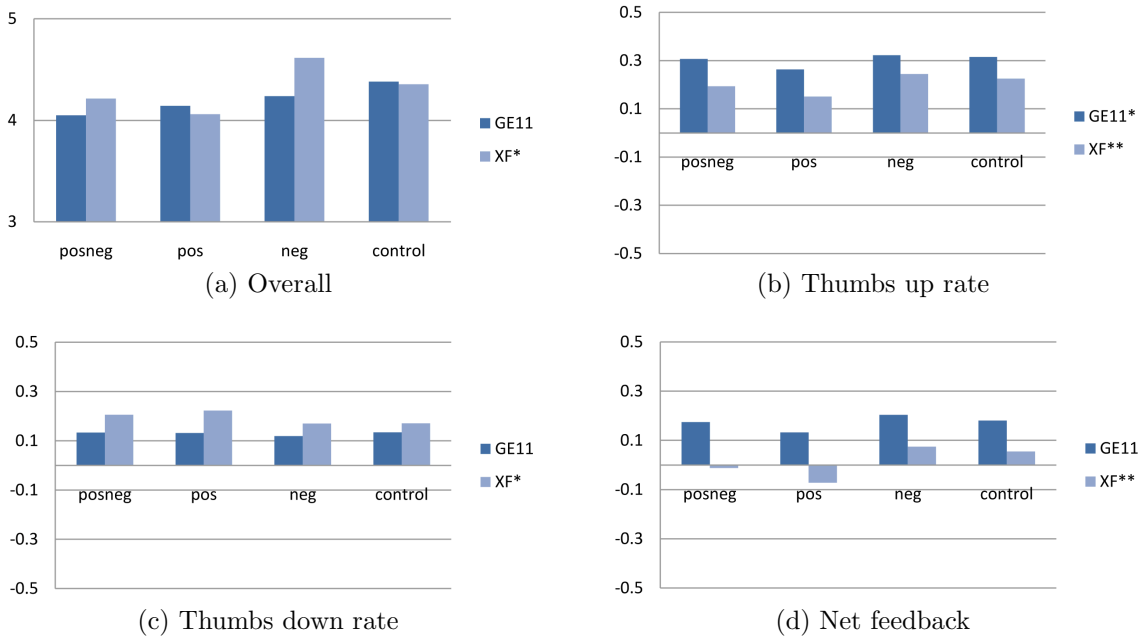
we present results from the topics alongside each other. We first look at algorithm-level sentiment, followed by document sentiment, and finally participant prior sentiment, with reference to other participant profile attributes throughout.

6.3.1 Algorithm Sentiment

Table 6.6 and Figure 6.2 show the mean feedback for the sentiment filtering algorithms. The average overall ratings for the sentiment filtering algorithms were slightly better than the midpoint of the 7-point scale, ranging between 4.06 and 4.61 for the X Factor and 4.05 and 4.38 for the Leaders' Debate. In each, the streams that upweight positive documents (`posneg` and `pos`) receive lower ratings than those that do not(`neg` and `control`). However, this difference in distributions is only significant for the X Factor ($p < 0.001$).

The feedback for the Leaders' Debate was far more positive with a thumbs up rate more than twice that of the thumbs down rate, whereas for the X Factor, the thumbs up rate was similar to the thumbs down rate. For thumbs up rate, 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 thumbs up rate ($p < 0.001$), X Factor thumbs down rate ($p < 0.05$) and X Factor $NetFeedback$ ($p < 0.001$). This



(a) Overall

(b) Thumbs up rate

(c) Thumbs down rate

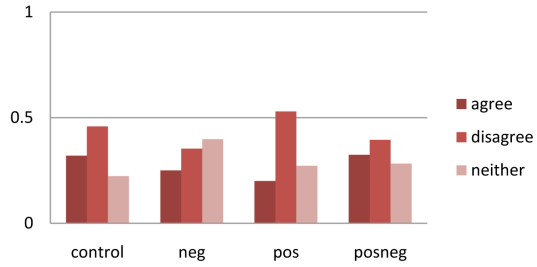
(d) Net feedback

Figure 6.2: Mean feedback for sentiment filtering algorithms

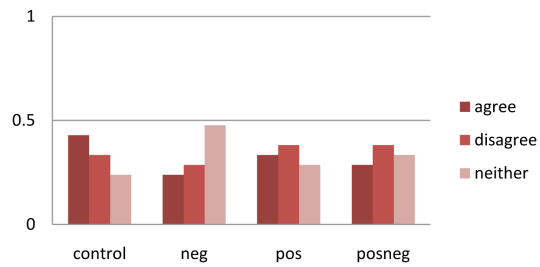
pattern is also present for *thumbs up* rate for the Leaders' Debate ($p < 0.05$). In both experiments, a document in the `pos` stream was considerably less likely to receive *thumbs up* feedback than in the `neg` or `control` stream. Also, the `posneg` algorithm performs significantly worse than the `control` for thumbs up rate and $NetFeedback$, although the effect size is smaller.

In Figure 6.3 we can see results of our stream-level survey feedback measures: *insightfulness*, *interestingness*, *informativeness* and *overall* rating. Although we do not observe any significant deviation in feedback between the different algorithms in either study, some interesting patterns emerge. Striking is the high percentage of stream feedback which agreed the streams were interesting, with algorithms averaging 59% for the X Factor and 74% for the Leaders' Debate. In both user studies, participants disagreed that the positive streams were insightful approximately half of the time. 53% of participants disagreed that the X Factor positive stream was informative, though this pattern does not appear to be present for the Leaders' Debate.

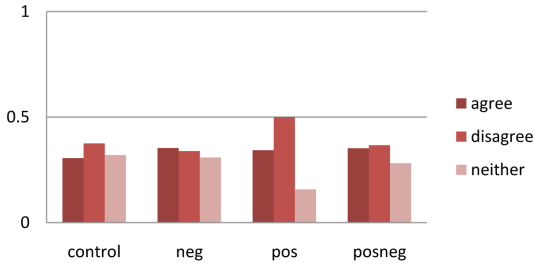
Table 6.7 contains the mean overall feedback for algorithms, broken down by participant attribute. We observe significant differences between the algorithm ratings for participants in different age groups and in different education groups for the X Factor study according to the between-subjects main effect ($p < 0.05$). The algorithm ratings



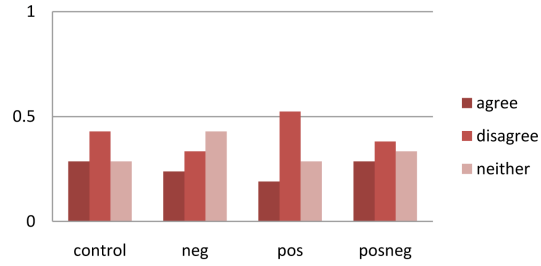
(a) X Factor: Informativeness



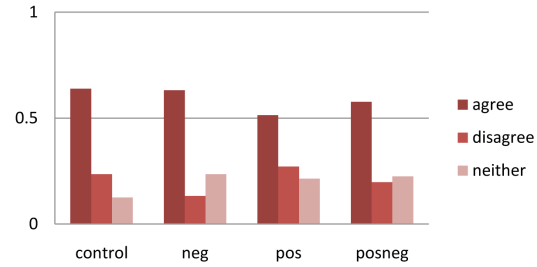
(b) GE11: Informativeness



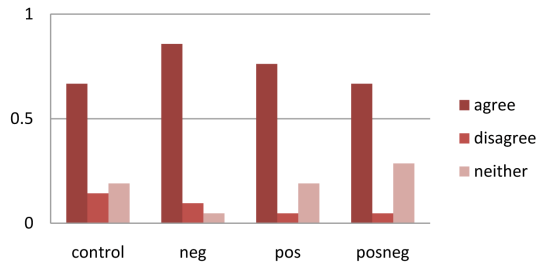
(c) X Factor: Insightfulness



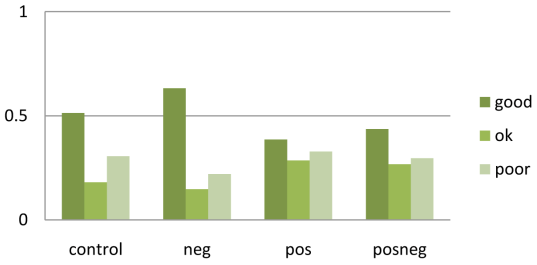
(d) GE11: Insightfulness



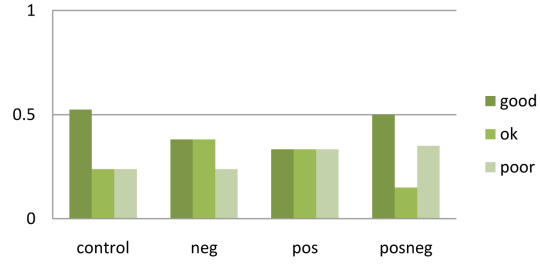
(e) X Factor: Interestingness



(f) GE11: Interestingness



(g) X Factor: Overall



(h) GE11: Overall

Figure 6.3: Overall and secondary feedback as categorical measures

			Sentiment Filtering Algorithm			
			posneg	pos	neg	control
Gender	XF	male	4.17	3.76	4.66	4.2
		female	4.29	4.63	4.52	4.65
	GE11	male	4.1	4	4.18	4.27
		female	4	4.3	4.3	4.5
Task Familiarity	XF	unfamiliar	3.84	3.86	4.32	4.09
		familiar	4.66	4.3	4.97	4.67
	GE11	unfamiliar	3.89	3.7	3.8	4.3
		familiar	4.18	4.55	4.64	4.45
Age	XF*	< 25	4.67	4.65	5.06	4.71
		≥ 25	3.74	3.43	4.15	3.98
	GE11	< 25	4.5	4	5.33	5
		≥ 25	4	4.17	4.06	4.28
Education	XF*	no degree	4.62	4.57	5	4.66
		degree	3.83	3.57	4.25	4.06
	GE11	no degree	4	4	4.67	4.67
		degree	4.06	4.17	4.17	4.33
Prior Sentiment	XF	positive	4.67	4.61	4.8	4.76
		negative	4.03	3.31	4.31	3.39
		neutral	3.79	3.93	4.36	4.75
		unfamiliar	3.33	3.67	5.17	4.17
	GE11	positive	4	3.67	5	5.33
		negative	4.06	4.22	4.11	4.22
		neutral	4.11	4.33	4.44	4.78
		unfamiliar	4.5	3.75	3.75	3.75

Table 6.7: Mean overall participant stream ratings, grouped by profile attribute and sentiment filtering algorithm (Significant differences noted for each attribute according to between-subjects main effect using the general linear model)

			Sentiment Filtering Algorithm			
			posneg	pos	neg	control
Gender	XF	male	-0.03	-0.1	0.06	0.02
		female	0.02	-0.03	0.09	0.11
	GE11	male	0.13	0.11	0.17	0.15
		female	0.23	0.15	0.24	0.21
Task Familiarity	XF	unfamiliar	-0.04	-0.07	0.03	0.03
		familiar	0.02	-0.07	0.12	0.09
	GE11	unfamiliar	0.13	0.10	0.22	0.19
		familiar	0.21	0.15	0.19	0.18
Age	XF*	< 25	0.06	0.03	0.13	0.13
		≥ 25	-0.08	-0.18	0.01	-0.02
	GE11	< 25	0.12	0.12	0.20	0.17
		≥ 25	0.18	0.13	0.20	0.18
Education	XF*	no degree	0.05	0.02	0.13	0.12
		degree	-0.07	-0.16	0.02	-0.01
	GE11	no degree	0.24	0.14	0.16	0.20
		degree	0.16	0.13	0.21	0.18
Prior Sentiment	XF	positive	0.02	0.01	0.07	0.08
		negative	-0.02	-0.14	0.08	0.02
		neutral	-0.06	-0.13	0.1	0.06
		unfamiliar	-0.05	-0.16	0.02	0.01
	GE11	positive	0.13	0.12	0.15	0.19
		negative	0.13	0.09	0.18	0.07
		neutral	0.26	0.18	0.25	0.25
		unfamiliar	0.06	0.08	0.17	0.13

Table 6.8: Mean participant stream net feedback, grouped by profile attribute and sentiment filtering algorithm (Significant differences noted for each attribute according to between-subjects main effect using the general linear model)

were far higher for participants under the age of 25 and for those participants who had a lower level of education. It should be noted that these two distinctions partitioned the user group similarly, but not identically. In Table 6.8 we see this same pattern is observed for $NetFeedback$ ($p < 0.05$).

Neither gender nor task familiarity demonstrate any significant different pattern for the different algorithms in terms of between subjects effect. In general however, we do observe a more positive response to the X Factor streams from younger and female participants, and from those who were already familiar with microblog search. This pattern is also observed in the participants for the Leaders' Debate. $NetFeedback$ confirms this observation although, as with overall ratings, no statistically significant between-subjects effect is observed for other participant attributes with respect to the sentiment filtering algorithms, beyond age and education for the X Factor.

6.3.2 Document Sentiment

Next we turn our attention to document-level sentiment. In the following results, document sentiment is the sentiment assigned to documents by the sentiment classifier. Each document may have either *positive*, *negative* or *neutral* sentiment, and will have received an annotation of *thumbs up*, *thumbs down* or *no feedback* from participants. We use the term "document" to refer to a single document presented to a participant, but this is more accurately a $< Document, Participant >$ pair, as the same document will have been presented to multiple users.

In Table 6.9 we see the contingency tables for *thumbs up* annotations with respect to document sentiment. For the X Factor, we observe a significant dependency between thumbs up feedback and positive documents and thumbs up feedback and negative documents ($p < 0.001$). Negative documents were twice as likely to receive a thumbs up from participants. Positive documents on the other hand were just half as likely to receive a thumbs up from participants as a neutral or negative document. For the Leaders' Debate we see no statistically significant sentiment-feeback dependencies and smaller effect sizes.

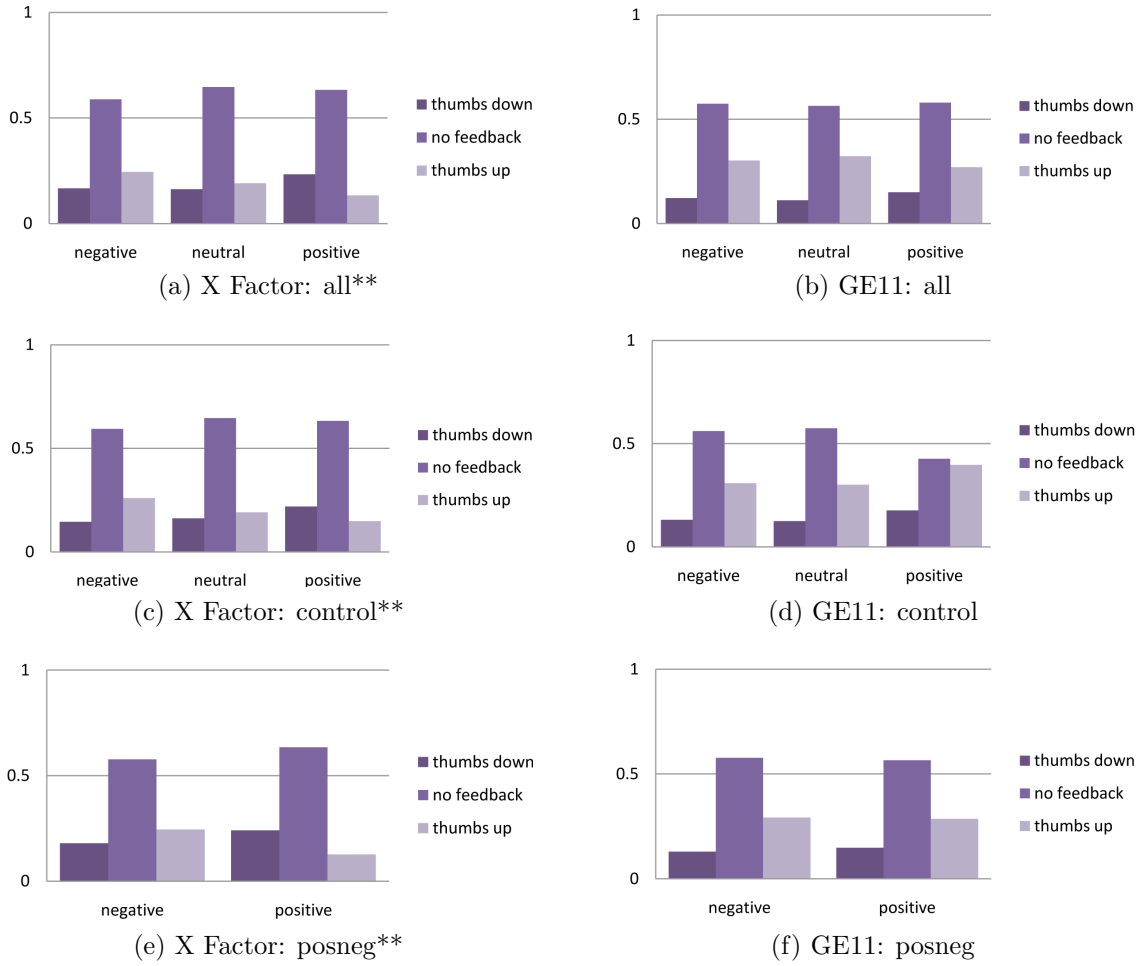
Table 6.10 contains contingency tables, this time for *thumbs down* feedback with respect to document sentiment. We again see significant patterns for positive and negative

			thumbs up		log odds
			yes	no	
XF	positive**	yes	1,098	2,435	-0.31
		no	7,145	7,760	
	negative**	yes	2,209	1,324	0.3
		no	6,813	8,092	
	neutral	yes	208	3,325	0
		no	881	14,024	
GE11	positive	yes	191	515	-0.08
		no	555	1,242	
	negative	yes	382	880	0.02
		no	364	877	
	neutral	yes	173	362	0.07
		no	573	1,395	

Table 6.9: Contingency tables with log odds ratio for thumbs up feedback per document sentiment type with significance according to chi-square

			thumbs down		log odds
			yes	no	
XF	positive**	yes	1,924	1,702	0.18
		no	6,319	8,493	
	negative**	yes	1,507	2,119	-0.16
		no	7,515	7,297	
	neutral*	yes	177	3,449	-0.11
		no	912	13,900	
GE11	positive*	yes	106	600	0.11
		no	215	1,582	
	negative	yes	155	1,107	-0.04
		no	166	1,075	
	neutral	yes	60	475	-0.08
		no	261	1,707	

Table 6.10: Contingency tables with log odds ratio for thumbs down feedback per document sentiment type with significance according to chi-square



(a) X Factor: all**

(b) GE11: all

(c) X Factor: control**

(d) GE11: control

(e) X Factor: posneg**

(f) GE11: posneg

Figure 6.4: Per-algorithm document-level feedback distributions

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 were 30% more likely to receive thumbs down feedback in our sample ($p < 0.05$). We also observed significant associations for neutral X Factor documents, which were 28% more likely to receive a thumbs down ($p < 0.05$). As with thumbs up, in general effect sizes are smaller for the Leaders' Debate than for the X Factor.

In Figure 6.4 we can see that the inverse thumbs down/thumbs up pattern for positive and negative documents in the X Factor is consistent across both the `posneg` and `control` algorithms. For the Leaders' Debate, the significant increase in thumbs up feedback for

positive documents appears to be due to feedback received during the `control` algorithm.

Examining the effect of these sentiment-feedback relationships further, we look at the log odds of thumbs up and thumbs down per document sentiment, given certain participant attributes (see Table 6.11). Throughout this data we see many significant associtions between feedback and profile attributes. From a sentiment point of view, we are interested in situations where this relationship varies with respect to different document types. To allow for easy comparison, we also present this data visually in Figure 6.7 with greener areas indicating a more positive association, and redder areas indicating a more negative association.

For age, education and gender, there is a sizeable significant effect for X Factor feedback. This is particularly the case for thumbs up. Thumbs up was significantly less than expected where participants (i) were aged 25 or older, (ii) were male and (iii) had a higher level of education ($p < 0.001$). The inverse pattern is observed for thumbs down and for both it is positive documents for which the greatest effect is observed.

We do not observe the same level of association for the Leaders' Debate data, with fewer significant differences, and smaller effect sizes. There are two anomolous results; males were just 53% as likely as female participants to thumbs up a neutral document ($p < 0.001$). It is possible that males are less likely to see neutral content as augmenting their viewing experience, or they simply do not place as high a value on the information contained in neutral documents. Those who were familiar with microblog search were 80% more likely to thumbs up a neutral document ($p < 0.05$). Perhaps this is due to a higher level of acceptance or trust of neutral microblog content from those that were familiar with consuming content in microblog streams.

6.3.3 Participant Sentiment

Referring to some of the aforementioned tables, we now examine results with respect to participant prior sentiment. As with much of the other profile attributes we mention previously, we observe no significant between-subjects effect for any of the prior participant sentiment categories with respect to different sentiment filtering algorithms. This holds for both overall stream feedback (Table 6.7) and for $NetFeedback$ (Table 6.8). There is little

Doc	Participants	X Factor		GE11	
		Thumbs up	Thumbs Down	Thumbs up	Thumbs Down
All	≥25	-0.26**	0.19**	0.02	-0.04
	Male	-0.14**	-0.03	-0.13**	0
	Degree	-0.22**	0.15**	0.04	0.17*
	Familiar with Task	0.05*	-0.1**	0.03	-0.01
	Prior Positive	0.07	-0.13**	-0.23**	-0.37**
	Prior Negative	-0.09**	-0.01	-0.05	0.14*
	Prior Neutral	0.06*	0.15**	0.3**	0.17**
	Prior Unfamiliar	-0.11**	0.06*	-0.66**	-0.12
Positive	≥25	-0.32**	0.26**	0.04	-0.02
	Male	-0.17**	0.01	-0.06	0
	Degree	-0.28**	0.25**	0.11	0.06
	Familiar with Task	0.04	-0.05*	-0.11	0.05
	Prior Positive	0.19**	-0.17**	-0.3*	-0.38*
	Prior Negative	-0.15**	0.05	-0.1	0.16
	Prior Neutral	-0.05	0.15**	0.3**	0.06
	Prior Unfamiliar	-0.14*	0.06	-0.13	0.02
Negative	≥25	-0.22**	0.12**	0	-0.01
	Male	-0.11**	-0.07*	-0.1	0.01
	Degree	-0.18**	0.06*	-0.01	0.19
	Familiar with Task	0.05*	-0.18**	0.01	-0.02
	Prior Positive	0	-0.1*	-0.25**	-0.48**
	Prior Negative	-0.04	-0.07*	-0.04	0.11
	Prior Neutral	0.09**	0.16**	0.31**	0.26**
	Prior Unfamiliar	-0.11*	0.09*	-0.24*	-0.19
Neutral	≥25	-0.24**	0.08	0.04	-0.16
	Male	-0.29**	-0.18*	-0.28**	-0.01
	Degree	-0.17*	0.04	0.09	0.32
	Familiar with Task	0.05	0.02	0.24*	-0.11
	Prior Positive	0.04	0	-0.11	-0.15
	Prior Negative	-0.3**	-0.14	0.02	0.13
	Prior Neutral	0.22*	0.14	0.27*	0.12
	Prior Unfamiliar	-0.08	-0.03	-0.34*	-0.18

Table 6.11: Effect size as log odds ratios for feedback type with respect to participant attributes (Significance according to chi-square)

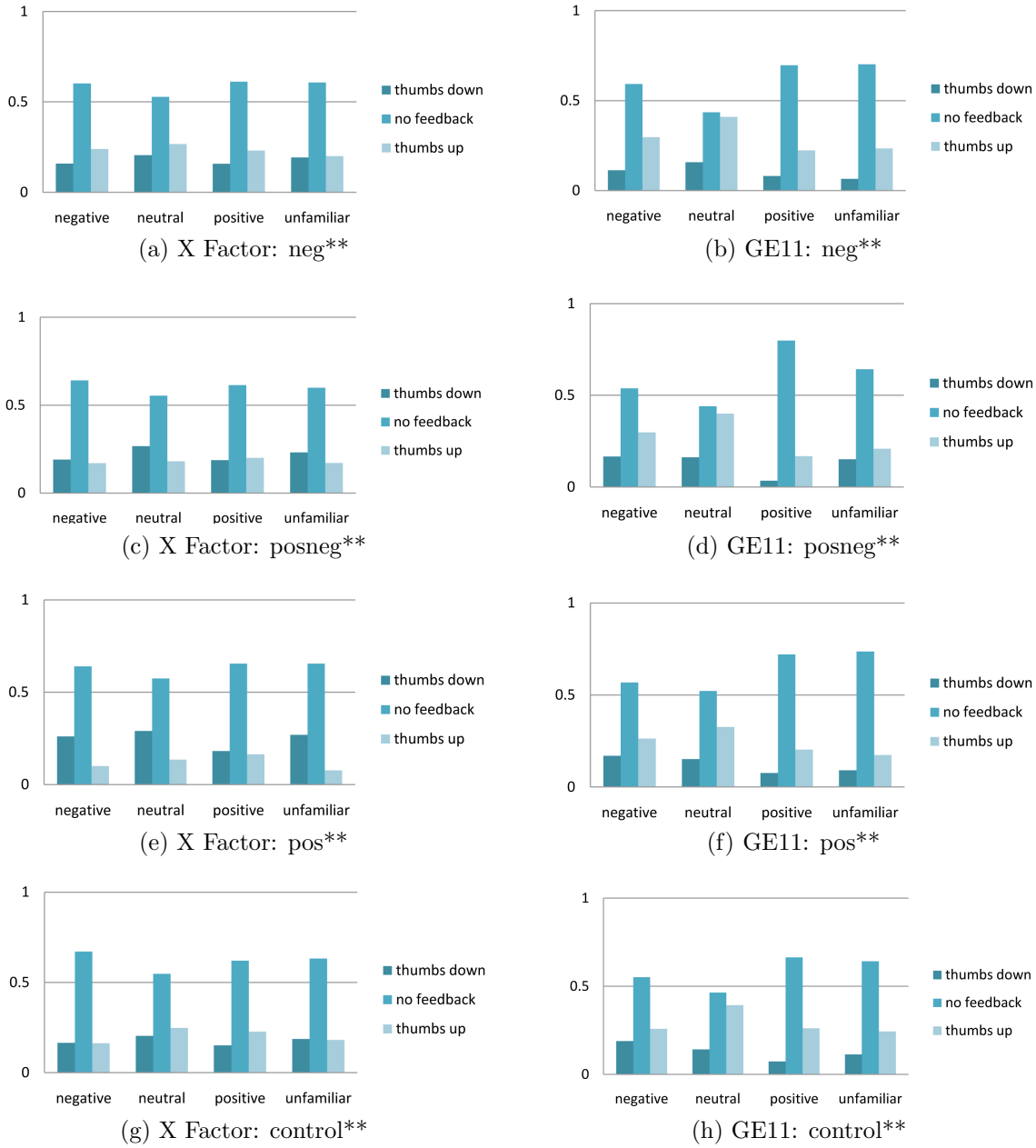
in either type of feedback that supports the notion that participants preferred algorithms which were aligned with their own prior sentiment.

For document-level sentiment, the difference between participants of different prior sentiment is significant for each algorithm in both studies ($p < 0.001$, see Figure 6.5). For the Leaders' Debate, we observe approximately the same distribution across the filtering algorithms. For the X Factor however, there is a noticeable difference for the `pos` algorithm, where participants who described themselves as negative, neutral or unfamiliar provided predominantly negative document-level feedback, whereas those who had declared themselves positive did not.

When we look at secondary metrics, we see significant differences between users of differing prior sentiment in how they describe the streams (see Figure 6.6). For both the X Factor and the Leaders' Debate, survey feedback for overall ($p < 0.001$) and insightfulness ($p < 0.05$) was varied across participants of different prior sentiment. For both topics, positive participants were most likely to give an overall "good" rating. Also in both studies, the difference in insightfulness appears to be due to the relatively low proportion of negative participants who agreed that the streams were insightful, and the relatively high degree of unfamiliar participants who agreed that the streams were insightful. Perhaps unsurprisingly, for both the Leaders' Debate and the X Factor, participants who had described themselves as positive rated the streams good more than half the time, far more than any other group of participant.

In the Leaders' Debate, participants who described themselves as negative were far less likely than any of the other sentiment categories to describe streams as interesting ($p < 0.05$). Although negative participants were also least likely to agree that the X Factor streams were interesting, this pattern is not statistically significant. For informativenes, we see little difference between prior sentiemnt for the Leaders' Debate. For the X Factor however, negative participants, and to a lesser extent positive participants, disagreed that the streams were informative, yet those who were unfamiliar agreed that they were informative more than 50% of the time.

Returning to the log odds figures in Table 6.11 and Figure 6.7, we see some interesting effects. 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.



(a) X Factor: neg**

(b) GE11: neg**

(c) X Factor: posneg**

(d) GE11: posneg**

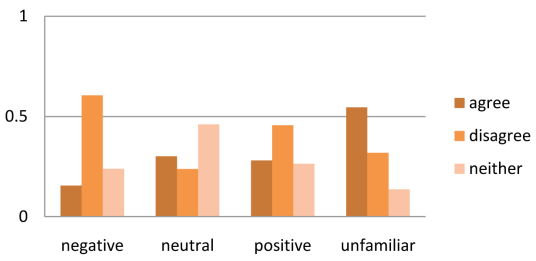
(e) X Factor: pos**

(f) GE11: pos**

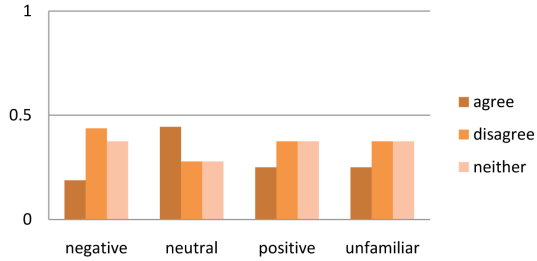
(g) X Factor: control**

(h) GE11: control**

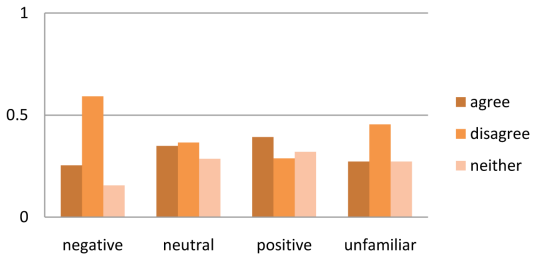
Figure 6.5: Document-level feedback distributions are different for groups of users with differing prior participant sentiment ($p < 0.001$)



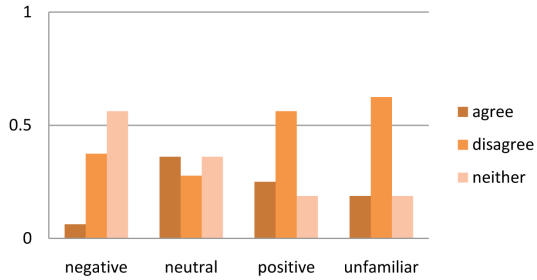
(a) X Factor: Informative**



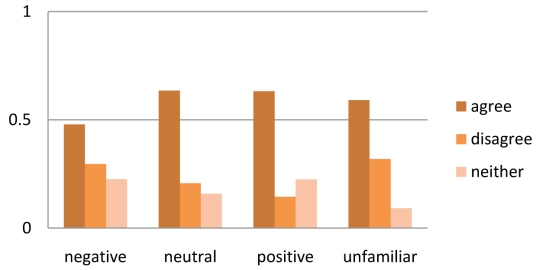
(b) GE11: Informative



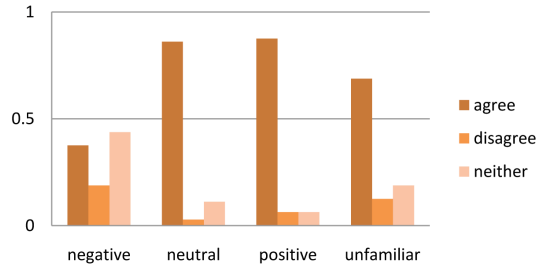
(c) X Factor: Insightful*



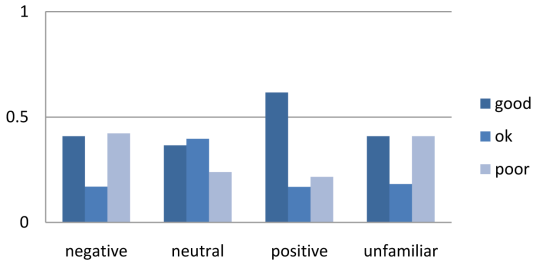
(d) GE11: Insightful*



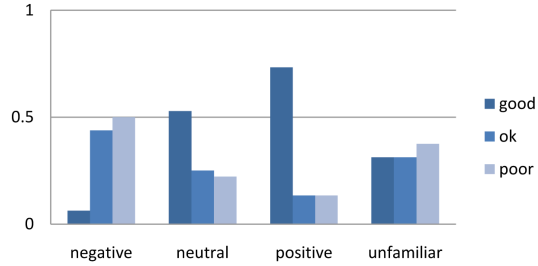
(e) X Factor: Interesting



(f) GE11: Interesting*



(g) X Factor: Overall**



(h) GE11: Overall**

Figure 6.6: Secondary feedback for prior particicpant sentiment

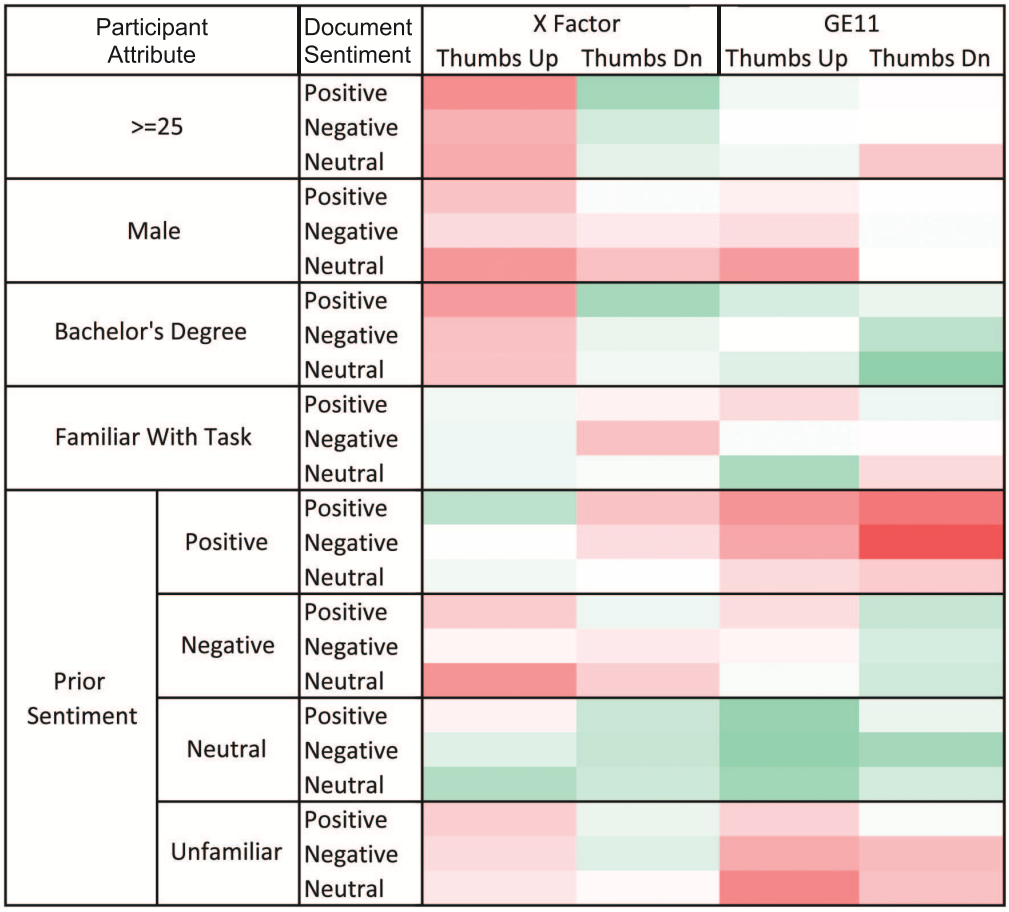
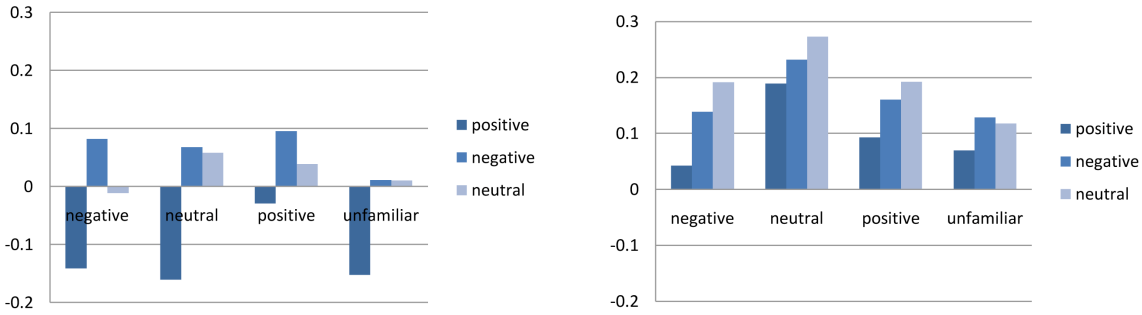


Figure 6.7: Visualisation of associations between user attributes and document feedback for different sentiment (Negative associations are redder, positive associations are greener)



(a) X Factor (b) GE11

Figure 6.8: Mean *NetFeedback* for prior sentiment groups

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 that 50% more likely to thumbs up a document regardless of sentiment. Overall neutral participants were more likely than others to offer feedback to the system.

The effect sizes observed for the X Factor are smaller, though in this case we do see 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. Negative participants were half as likely to thumbs up a neutral document yet neutral participants were 67% more likely to thumbs up a neutral document. The *NetFeedback* scores for document sentiment types, grouped by participant prior sentiment can be seen in Figure 6.8. The patterns for the X Factor and the Leaders Debate are quite different, although positive documents are consistently perceived the worst in each grouping.

6.4 Discussion

In our user studies, we have captured and analysed a substantial amount of data and we now discuss this analysis with respect to our research questions.

Overall, there is little to suggest that employing a sentiment filtering algorithm on a real-time stream has a significant impact on the user experience during real-time microblog search for our chosen topics. We repeatedly see similar feedback distributions and patterns

for each of our algorithms. An exception to this is the `pos` algorithm, which produces a number of poor results. This is perhaps to do with participant dissatisfaction with positive content, or perhaps the absence of other types of sentiment. The `neg` algorithm performs similarly to the `control` stream despite having much less neutral and positive content. We speculate that this negative content is the content that is valuable to the searcher.

In either case, the `control` proves to be a strong baseline and it is unlikely with our current approach that any of our algorithms would outperform the `control` algorithm. It appears that the effect of altering sentiment in the stream is minimal, and certainly does not augment the experience for users in general, or for any particular user group.

Once we begin to examine feedback at a document level however, we begin to see more significant patterns. Across both experiments, we see positive documents are negatively received by participants and negative documents are positively received, reinforcing what we see at algorithm level. But what is so attractive about the negative documents, and what is so jarring about the positive documents? From qualitatively examining the data, we speculate there are two effects at work. First, the documents classified as positive tend to be those where the sentiment is explicit and stated in simple terms, and thus easier documents for the classifier to identify. These documents are frequently just a few words stating support for a topic entity, offering little in the way of content for the searcher. This is possibly reflected in the high proportion of participants who disagreed that the positive stream was insightful for both topics. Secondly, we found the majority of humorous and critical content is negative in nature. Perhaps this type of content is successfully being identified as negative by the classifier, and is considered valuable by the user.

For the X Factor, many of the significant differences we observe are perhaps to be expected due to the target demographic of the show. We observed a division between old and young participants, and between male and female participants. We saw that this difference was measureable with respect to document-level feedback. This effect varies as expected with document sentiment type, though this difference is relatively small. The Leaders' Debate has arguably a wider relevance, and thus a less targeted audience. It is perhaps for this reason the demographic effect sizes observed are much smaller than that for the X Factor with fewer significant associations.

The different responses from participants of different prior sentiment is intriguing.

Clearly, the participant perception of insightfulness, informativeness and interestingness is contingent on their own beliefs and preconceptions. On the whole, participants rated the content highly for interestingness, but were more split for insightfulness and informativeness. What is more interesting, is that these ratings consistently varied between groups of participants with differing prior sentiment. Note too, that it is not just positively and negatively disposed participants, but also unfamiliar and neutral participants who each display unique behaviour. Although it is difficult to discern a common pattern, there are a number of large effects observed for the Leaders' Debate with respect to feedback. For the most part, these effects are consistent across document sentiment types, and appear to be more different general approaches to the task from participants with disparate prior sentiment. In the debate, the nature of the subject matter is controversial, serious and impactful, and perhaps prior sentiment has a stronger bearing on such content than is the case for the X Factor, where the subject matter is, on the whole, more light-hearted and less consequential, and the effect sizes smaller.

Regarding user demographic profile attributes, our studies reveal a number of significant associations. On reflection there appear to be two reasons for this. During the X Factor experiment our younger, female participants gave much more positive feedback than others, pointing to a relationship with the target audience for the show. More generally, it appears that while participants' profiles had an observable effect on their task performance, this was mostly independent of document-level or stream-level sentiment. It seems in fact to be more related to their general approach to microblog search.

6.5 The GE11 Twitter Tracker: Monitoring Public Political Sentiment

As discussed in Chapters 2 and 4, sentiment analysis offers many opportuities for mass-opinion measurement. During the General Election we developed the "GE11 Twitter Tracker" in collaboration with our partner, an Irish news website[2]. This system expands upon the data and tehniques used in our experiments to produce a microblog analytics

[2] `http://www.thejournal.ie`

Figure 6.9: The GE11 Twitter Tracker: Sentiment Series

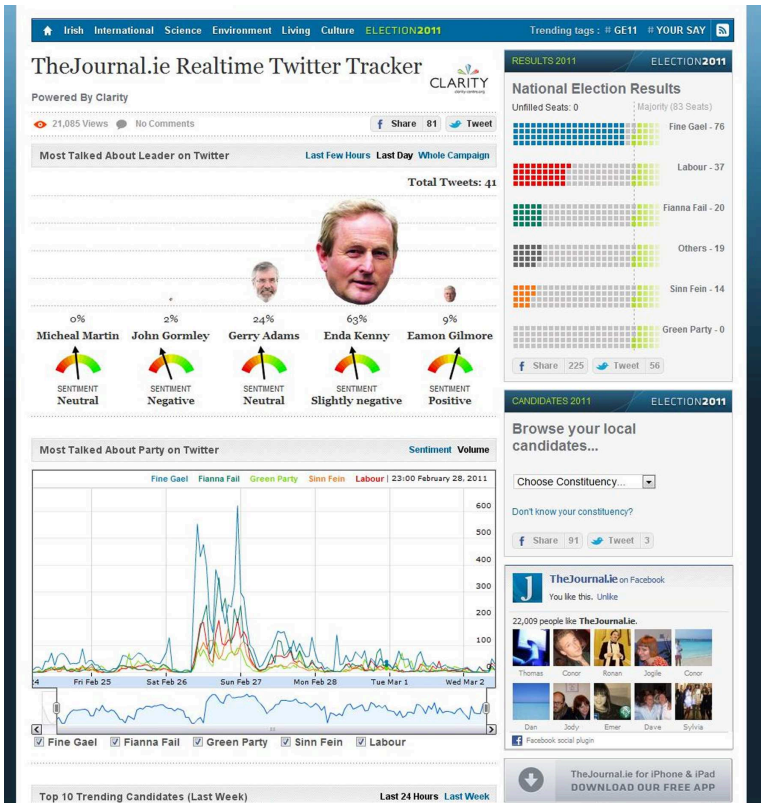


Figure 6.10: The GE11 Twitter Tracker: Volume Series

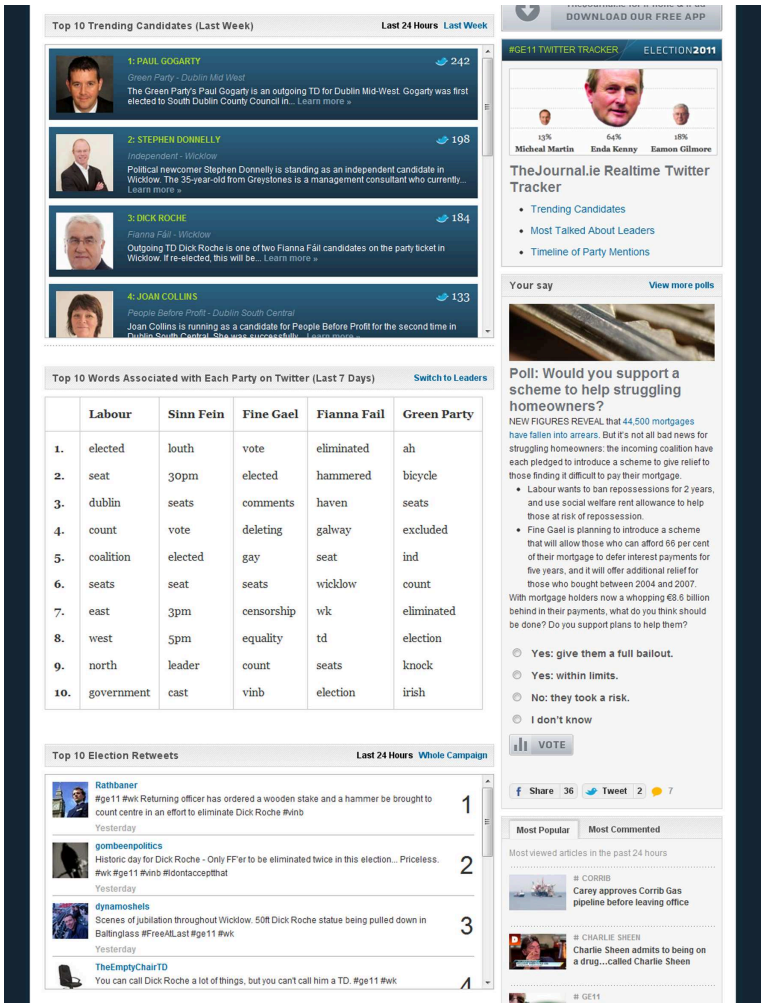


Figure 6.11: The GE11 Twitter Tracker: Trending Candidates, Associated Terms and Top Retweets

service. The purpose of the "GE11 Twitter Tracker" was to allow users, and our partner's journalists, to tap into the content on Twitter pertaining to the election, through an accessible dashboard-style interface. To that end, the Twitter Tracker featured a number of abstractive and extractive summarization approaches as well as a visualisation of volume and sentiment over time (see Figures 6.9, 6.10 and 6.11).

The General Election had attracted many of the media, electorate and candidates to Twitter, who had no previous experience of microblogging. Tracking a real-time topic on Twitter can be an involved process, meaning monitoring a stream of documents over an extended time period, or checking the stream frequently. The vision for the Twitter Tracker was to provide at-a-glance summarization of activity on Twitter for the casual microblog user, while also providing more in-depth analysis for power users.

The features of the Twitter Tracker were as follows:

- *Party Leader Volume*: Volume of tweets relevant to party leader over time, expressed as a percentage. The volume of the relevant data was also visually represented by scaling the size of each leader's photograph in line with their relative volume.

- *Party Leader Sentiment*: Sentiment of each party leader desccribed on a 7-point scale: *very negative*, *negative*, *slightly negative*, *neutral*, *slightly positive*, *positive*, *very positive*. Sentiment was also visually indicated on temperature gauges.

- *Party Volume*: Number of tweets relevant to parties were graphed over time. This data could be annotated by our partner's journalists with links to news stories.

- *Party Sentiment*: Party sentiment graphed over time.

- *Trending Candidates*: Using the Twitter metaphor of "trending topics", we tracked mentions of the 566 candidates for the election. We then displayed the top ten highest ranked in terms of volume of relevant tweets.

- *Associated Terms*: Using TF-IDF we identify terms most associated with (i) parties and (ii) party leaders.

- *Retweet Charts*: We display a top ten list of the most reweeted tweets.

For many of the features, users could select a number of different time periods such as "last 24 hours" or "whole election" to allow a flexible temporal granularity. For the sentiment measures we used the following formulation of sentiment, a log ratio of the volume of positive and negative sentiment for a topic, x:

$$Sent(x) = \log_{10} \frac{|Pos(x)| + 1}{|Neg(x)| + 1} \tag{6.2}$$

The classifier used was the same as that used in the user studies.

The Twitter Tracker was a resounding success, receiving approximately 1,000 pageviews per day during the election. It was featured on national television (Tonight with Vincent Browne, TV3), radio (The Right Hook, Newstalk FM), technology news website, Silicon Republic (Kennedy, 2011), and was acclaimed by the Nieman Journalism Lab in Harvard University (Kelly, 2011). At time of writing, a retrospective version of the GE11 Twitter Tracker is still live[3].

The Twitter Tracker was one of the first public systems of its type. Throughout the election, it demonstrated the power of using content analysis, and in particular sentiment analysis, of microblogs to drive real world, real-time analytics applications.

6.6 Conclusion

In this, our final experimental chapter, we have described our user studies to examine the role of 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 studies with the Channel S system. We detailed our methodology including sentiment analysis configuration, ethical considerations and experimental set-up.

Our results show that altering the sentiment in a stream results in little difference in feedback, with the exception of upweighting positive documents which can attract more negative feedback. At a document level, we observed a similar pattern where positive documents are more likely to receive negative feedback, and negative documents are more likely to receive positive feedback. We consistently see different feedback to the system from users with differing prior sentiment. Similarly, participant demographic profiles ap-

[3] `http://www.thejournal.ie/twitter-tracker`

pear to account more for the general approach of participants to the task, rather than for any particular sentiment-related aspect. Perhaps as real-time search becomes better understood, we will gain a clearer understanding of the motivations and approaches of different users. For the X Factor we do see some patterns with respect to sentiment and user profiles, possibly due to the show having a well-defined target demographic.

Although the effects we see are mixed, it is clear that sentiment in a real-time search system is a measurable quantity. It is also clear that in many circumstances 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 real-time automated sentiment analysis, and its use in such systems.

Chapter 7

Conclusions

In this, the concluding chapter of the thesis, we summarise our work, conclusions and contributions. We have explored in depth how sentiment manifests itself in microblog posts and real-time microblog search scenarios. As a relatively new area of research, much of our work has been progressive and as such, deserves suitable reflection and contemplation. As well as summarising our research and conclusions, we also reflect on the context of our research, future directions and how our methodologies might be improved.

We first review the content of the thesis in Section 7.1. This is followed in Section 7.2 by a discussion of our conclusions with respect to our hypotheses and research questions. In Section 7.3 we reflect on our work and outline directions for future research.

7.1 Summary

In our introductory chapter (Chapter 1) we introduced the concept of microblogging. We discussed the high impact that microblogging is having on today's world and how this has motivated the need for efficient real-time search systems. We also introduced sentiment analysis and described how, despite its maturity as a field, it has struggled to gain credence for use in search systems. Following from this, we motivated our decision to investigate the utility of using sentiment to augment real-time microblog search. This was formalised in our hypotheses and research questions.

In our overview chapter (Chapter 2), we reviewed the state of the art in sentiment analysis. We also reviewed information retrieval research literature, with a particular

focus on information filtering, and the new field of microblog search. We detailed the evaluation methodologies that are used in sentiment analysis and information retrieval and how this pertains to our research, including a comparison of user-centric evaluation and static corpus evaluations. We also introduced the notion of different sentiment levels — document-level, user-level and stream-level — which we use throughout our work.

We then continued to introduce our experimental system, Channel S, which we use to conduct our user studies (Chapter 3). We described its architecture, design and implementation, and explained how it supports our evaluations. Following from this, we introduced our experiments. Experiments I and II examine subsections of the system, while Experiment III is a full deployment of the system with real users, in real-time.

Our first experimental chapter (Chapter 4) concerned the task of using supervised learning to identify sentiment in microblog posts. We reviewed the literature regarding sentiment classification in user-generated content, with a particular focus on supervised learning. We developed a corpus of microblog topics and labelled posts. We evaluated a number of different feature sets and drew comparison with supervised sentiment classification in three other domains: reviews, microreviews and blogs. Our results showed a favourable classification accuracy for microblogs, though we struggled to rival the accuracy achieved in review-style data. We found it difficult to improve upon a baseline accuracy which uses unigram vector document representations.

For our second experiment (Chapter 5), we chose a subset of our topics, and used our labelled data to simulate a real-time search task with 16 users. This had the advantage of using high-accuracy manually labelled data, rather than relying on potentially noisy automated sentiment analysis. In testing simulated scenarios, we were also able to cover a variety of topics. Indeed we found a number of interesting patterns among the topics, including a correlation of positive sentiment with subjectivity; the more subjective content about a topic was, the more likely that content was to be positive. The opposite appeared to be true, however, if we just consider topics which were companies. Only in the minority of cases did we observe a significant difference in user feedback with respect to different filtering algorithms. We did, however, observe statistically significant associations between user profiles and feedback, user prior sentiment and feedback and between document-level sentiment and user feedback.

Lastly, we deployed the full Channel S system for three live events: two shows of the popular television programme, the X Factor, and the Leaders' Debate during the Irish General Election, 2011 (Chapter 6). With a larger sample of users, approximately 20 per event, we were able to capture a sizable amount of real-time feedback data, using three sentiment filtering algorithms and a control algorithm to present selected microblog posts to participants. We discussed the ethical considerations around these studies and detailed our ethics materials. We also explored the necessary sentiment analysis configuration and how we overcame the dearth of positive content relevant to the election using a boosted classifier. For the X Factor, we observed significantly different reponses for our algorithms, most prominently the negative reaction to positive streams. This effect is also present for the debate, but we only see it for explicit thumbs up feedback. We looked at prior sentiment and user profiles more deeply than we had previously in the simulated real-time experiment. We found that demographics seem only in some cases to be associated with sentiment algorithms. We saw how in general, though, user prior sentiment and demographic has a strong bearing on how a participant approaches the task. We also explored how these factors are related to document-level sentiment.

In Chapter 6, we also described the GE11 Twitter Tracker, a system we developed for monitoring public political sentiment in conjunction with an Irish news website. We described the vision of allowing novice users to understand activity on Twitter at-a-glance, as well as allowing more inquisitive users the opportunity to explore the analytics further. We described how sentiment analysis was used, in conjunction with other content analysis and frequency-based measures, to provide dashboard-style analytics. The system also allowed for journalists to annotate the analytics with their stories. The system was successful, receiving a substantial amount of pageviews, and considerable media attention.

In total, we have run a series of three experiments, generating a large amount of participant feedback and associated analysis. Along the way, we have also developed a substantial corpus of data consisting of almost 20,000 labelled documents and 70 topics and sentiment targets. We have also developed experimental materials such as surveys, annotation guidelines and ethics documentation as well as a refined annotation tool for labelling documents for sentiment. Lastly, we have designed, architected and implemented a web-based system for evaluating real-time microblog contextual search with sentiment

filtering algorithms.

7.2 Conclusions

Recall from our introduction chapter that our work has two hypotheses. Our primary hypothesis states that sentiment in real-time streams has a significant impact on the perceived quality of the stream from the point of view of the user. Our secondary hypothesis concerns our ability to identify sentiment in a new textual domain, microblogs, using supervised learning techniques. Satisfying our secondary hypothesis is necessary to empircally support the evaluation of our primary hypothesis. In this section we summarise our conclusions, first towards the latter of these hypotheses, and then towards the former.

Our efforts to classify ad-hoc sentiment in microblogs have been successful. Our accuracies of 74.85% for binary positive-negative classification and 61.3% for three-way positive-negative-neutral classification each demonstrated the considerable ability of supervised classifiers to discriminate between textual content in microblog posts according to sentiment. Although this accuracy is lower than we observed for review-style texts, it is important to remember that the task of ad-hoc sentiment analysis is fundamentally a harder problem than review classification. Review classification benefits from limited and consistent domain vocabulary and semantics, homogenous topics and inherent subjectivity. Compounding this disparity is the difficulty in obtaining training data for microblogs; ad-hoc sentiment annotation tasks are more prone to problems of ambiguity and the documents are not annotated in any way by the author (as is often the case with review content).

Another concern we had before performing our sentiment analysis evaluation was that the nature of the microblog domain would prove troublesome for sentiment classification. As a short-form domain, sparse feature vectors could have proven difficult for the classifier — the mean document length in our sample was just less than 18 words, with many documents containing considerably fewer. We found that this was not in fact a problem as it appears that the short texts are less prone to the problems of topic and sentiment drift which plague classification in longer ad-hoc domains, such as blogs. It was interesting to observe that in review data which is less prone to topic and sentiment drift, the longer

texts were easier to classify than the shorter texts.

Microblogs, as with many other forms of user-generated content, contain a high degree of non-standard language and punctuation usage. We found these sociolinguistic features to be among the most discriminative; indeed we use a tailored tokeniser in our final experiment which specialises at extracting these features. We found that users are using emoticons and punctuation (for example) to add intonation and context to their posts. This echoes previous reseach, which finds that short-form CMC domains tend to be a hybrid of written and speech-like language. This type of text can be problematic for sophisticated linguistic feature extraction, and we found that a unigram baseline is strong and difficult to beat with alternate vector representations.

Overall, it is clear that we can satisfactorily classify sentiment in microblog posts. This conclusion is reinforced later in our experiments when we observe statistically significant differences in responses to different types of automatically classified sentiment, indicating that our sentiment analyser is successfully, and likely correctly, discriminating between content.

In order to disprove our primary hypothesis we would have to have seen no significant difference in perceived information quality from participants. In fact our experimental observations reveal many statistically significant differences. Throughout both our simulated and real-time experiments we see significant patterns developing in terms of user feedback and sentiment, confirming our intuition that sentiment does indeed impact real-time microblog search.

In our simulation experiment we found a number of patterns which demonstrate the role of sentiment in real-time search. First we found a significant relationship between document sentiment and document feedback, with participants preferring neutral documents and disliking positive documents. We saw that even the underlying sentiment distribution for a topic was indicative of how well the content would be perceived; users tended to dislike documents for topics with a high degree of subjectivity, rating them poorly for informativeness. Participants also rated the positive algorithm poorly when the topic itself was a topic that attracted a high degree of positive sentiment.

In our real-time experiments, we again observed the significant impact that sentiment has on real-time search. For the X Factor, we found that positive documents were neg-

atively received and negative documents received disproportionately more positive feedback. The data suggests a similar pattern for the Leaders' Debate (though in this case the pattern is only statistically significant for thumbs down for positive documents). In both experiments we see significantly more negative feedback for the positive streams and significant differences in variance across the sentiment filering algorithms. Specifically in the X Factor we saw how the different sentiment streams were perceived differently by different age and education groups. We saw a number of significant differences between user profile groupings but (with the exception of age/education for the X Factor) it is difficult to identify any pattern which is specifically contingent on sentiment. It would appear that patterns observed for user profile differences are largely due to the participants' approach to the search task and topic, and not necessarily associated with document or stream sentiment.

In both studies, we see a radical difference in feedback from users with different prior sentiment towards the event topic, with participants from each of our four prior sentiment categories (positive, negative, neutral and unfamiliar) exhibiting very different behaviour. The four categories of user give significantly different feedback in all of our feedback measures, throughout our experiments. Furthermore, there is evidence to suggest that a participant's prior sentiment it at least somewhat aligned with their stream preference. For example in our simulations we saw that positive participants rated the negative stream the lowest and negative participants rated the positive stream the lowest. We also saw negative participants rating the negative stream by far the highest. We observed significant differences between participants of differing prior sentiment in terms of the feedback they gave documents of different sentiment in our live studies. However, it appears that prior sentiment has a strong association with perceived microblog quality in general and this effect is much stronger than the associations with any particular type of document sentiment.

Despite these positive results for sentiment, there are some aspects of our system which have not demonstrated the effects we had hoped. For the most part, upweighting subjective documents (i.e. the `posneg` stream) demonstrated no observable effect in our study. We have not shown that this distribution of sentiment in microblog streams is perceptibly different from the control distribution. Also, 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, so it is possible the smaller sample size meant that the difference in feedback at a stream level was more difficult to detect than it had been for the X Factor. We do see some patterns for the debate, for example the positive stream receiving significantly less thumbs up feedback than the control, and on the whole positive documents receiving significantly more thumbs down feedback than expected. These results suggest that perhaps some of our inconclusive debate results are subject to type II error and, in a larger study with greater power, more similar significant patterns would emerge.

There are differences between the Leaders' Debate and the X Factor in terms of our observations. For example, one of our research questions concerns whether document sentiment is indicative of how that document might be perceived. In the X Factor and the simulated search scenarios this was very much the case, but this was not as evident for the Leader's Debate. We must therefore conclude that this type of effect, though it exists, is perhaps limited to certain topics and search scenarios.

In summary, we have successfully demonstrated that microblog posts can be classified according to sentiment and that this can be integrated into a real-time microblog search system and experimental framework. We have demonstrated several cases where filtering sentiment significantly alters the user's perception of the quality of the stream, usually detrimentally. We have also made a series of observations which demonstrate an association between user feedback and sentiment at a document level. Throughout, we considered results with respect to user demographic and found many statistically significant patterns. Similarly, we found that comparing feedback to user prior sentiment consistently reveals significant patterns. We conclude also however, that in some cases, these profile variables have a stronger association with task feedback in general, rather than any sentiment-specific aspect.

7.3 Reflections and Future Work

Our work has taken a journey through the world of real-time microblog search and sentiment analysis. We have learned a lot with respect to these areas, and are now in a position

to reason about the appropriateness of our assumptions and experimental methodology. In this section we discuss these aspects as we reflect on our work and present thoughts on future directions for research.

The definition of sentiment we used is well-founded in the literature. We assumed that all relevent content may be considered as one of positive, negative, neutral or mixed with respect to a topic. Furthermore, we assumed these were to be interpreted with respect to emotions, opinions, evaluation and speculation. These sentiment categories allow us to easily construct experiments as the categories fit neatly into a machine learning classification problem. They are however a trivialisation of sentiment as it exists in the real world. Is the sentiment we are trying to capture from the point of view of the author? Perhaps a cited source? Or perhaps we should consider sentiment from the point of view of the topic — how does the content reflect on the topic in general? Perhaps the most difficult distinction is between objective and subjective content. Is bad news about a topic objective? Or, if not subjective, containing sentiment of some kind? These are problems which in current systems are tackled through assumptions, and sometimes arbitrary distinctions. Perhaps as the field of sentiment analysis progresses and highly-focused applications are developed, sentiment can be defined more rigorously for specific scenarios and applications.

Another troublesome aspect is the prevalence of mixed sentiment content. It is promising for sentiment analysis technology to see such a low proportion of mixed sentiment in our microblog content (typically around 5%). In many experiments (including some of ours), mixed examples are excluded from training data as they are too ambiguously defined and prone to inconsistent labelling. This however does not mean that mixed content will not be present in real-world testing and it will therefore inevitably be classified incorrectly. It remains an exercise for future work to identify how these document can be better accommodated in this type of supervised learning framework.

Sentiment analysis is a specialisation of what might be called the wider *document understanding problem*. It is necessary to understand more than just the literal interpretation of content. There are also social and cultural components. In the use case where a user reads the microblog posts from those whom they follow, the user has an internal knowledge base of the author from previous content, as well as real-world observations

and interactions. This informs the context of the microblog reading task and helps users to distinguish for example between complex human linguistic concepts such as irony, humour and sarcasm. In our experiments, the profiles of the authors were not (by and large) known to the users, and thus participants took all content at face value, possibly missing such subtlety of language. Upon qualitatively examining our data, we found humour and wit were common to much of the highly-rated data. We speculated that, as a large portion of this humour is derisory, perhaps it is being inadvertantly isolated as negative content in our system. Further study is required to explicitly evaluate the role of humour and whether this can be accounted for in an automated fashion.

We have endeavoured in our research to first study a diverse range of topics, and then evaluate with respect to two real-time topics in much closer detail. We made a reasonable assumption that Twitter's trending topics are indicative of real-time information needs. A natural extension to our experiments would be to run them with respect to other events intended for a live audience such as television programmes, sports matches, arts and music performances and others. These topics by their nature are designed for a shared audience, and perhaps represent only a subset of those that we may consider. There are likely other topics worth considering in microblog search, however revelation of such topics may perhaps only happen with a study of microblog query logs. A large portion of the utility of real-time microblog systems concerns efficient dissemination of information, particularly related to breaking news topics. Microblog systems allow information to be channelled efficiently through a social graph in minutes, if not seconds. Our real-time system in its current incarnation does not support evaluation with respect to these types of topics. It is a future challenge for microblog search to develop methodologies which can account for more ad-hoc contexts for real-time microblog search evaluation using real-time feedback, perhaps through a distributed web-based system.

As with any supervised learning task, we had to invest significant resources in developing labelled data to train our classifiers. For our final experiments, we used specifically trained classifiers. We were satisfied that we had reached a ceiling of what can be accomplished with human annotators — moderate to good agreement, and an annotation rate of approximately six per minute. This is still a significant investment of time and human effort. We saw that a lexicon-derived approach did not perform comparably, although we

acknowledge that our lexicon-based classifier could likely be improved by incorporating, for example, topic proximity, coreference resolution or negation. There is however opportunity for research to pursue building a hybrid approach to classification, for example where data is used to modify lexicons to tackle the problems with topical domain transference. Alternatively, known data or sentiment dictionaries can be used to bootstrap the learning process in a semi-supervised approach. Another approach which we could have considered is implementing active learning to maximize the efficiency in choosing documents for annotation.

When we started out this research, our focus had been to evaluate sentiment filtering patterns, and perhaps to observe certain filtering algorithms which improve the quality of the streams. We were not able to achieve this, and in some cases this reduced the quality of the streams. In modifying the sentiment in the stream we are possibly introducing negative effects other than simply 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.

As our experiments continued, we realised that it made more sense to think about the content at a document level, rather than a stream level. Even from a methodological point of view, document-level feedback was easy to capture in a way that was intuitive and instinctive for participants, and which yielded a substantial amount of data. In contrast, stream-level feedback was difficult to capture in a large quantity and the low variance indicates that we were not able to capture the same intricacies that we could with document-level feedback. Notwithstanding this, we can be satisfied that our stream-level feedback is consistent with our document-level feedback and that our survey measures were helpful in discerning the motivations behind user feedback. Also, our adaption of the like and dislike social feedback metaphor to real-time search evaluation appears to have been successful, and is certainly a feedback mechanism that we would use again in a future system.

Our studies suggest some interesting patterns concerning user profiles and they certainly suggest that who a user is, and what they believe, has a fundamental effect on how they perceive content in real-time microblog search. In our experiments, we have recorded

user profiles and preferences, though we have not attempted to incorporate this into our system. Similarly, we have not used real-time feedback to inform our system of individual user preferences. Given our observations, a logical next step is to explore the potential for introducing a personalisation component to the system. Such a method could also use features from the underlying social graph to better model content of interest to users. Incorporating sentiment we could use the sentiment profiles of peers, groups and similar individuals to tailor the search system. Our demographic and prior sentiment observations offer a first investigation into this area, and research with larger sample groups and a degree of personalisation offers a promising future avenue for research.

Our experiments have been intrinsically tied to Twitter as a microblogging platform, and as such, are subject to the parameters of the Twitter system. Twitter has imposed constraints on the network, such as limiting content to text with a maximum length of 140 characters, and using a follower-following paradigm for their social graph. This is not however the only manifestation of microblog systems. There are several other pervasive microblogging platforms, each of which have different constraints that define the nature and culture of that particular platform. For example, there is the more private network of status updates in Facebook, and the platform specifically designed for enterprise microblogging, Yammer. We chose Twitter due to its availability and popularity, though future work should consider other platforms to give a more rounded perspective on the general nature of microblogging as a platform for publication, communication and search.

Finally, it is important to note that the techniques we develop are not just of interest to systems supporting information seeking. Real-time topic monitoring and sentiment analysis is of interest in an aggregated form as summarisation or visualisation, often in the commercial world referred to as "analytics". These can be used to support traditional search applications also. This type of information presentation and interaction is of obvious commercial value for parties who wish to monitor entities in which they have an interest. During the course of this work we have demonstrated the flexibility of the underlying technology to support a real-time analytics site with the public Twitter Tracker we developed during the Irish General Election. This aggregate analysis is also proving an excellent avenue of research for those who wish to use social signals to inform statistical models, such as financial forecasting, or polling.

As with much research, in answering our research questions, we have perhaps raised more questions for us and others in the research community to pursue. On the whole, we conclude that real-time microblog content and sentiment analysis provide an exciting new opportunity to study methods to develop real-time social information systems.

Appendices

Appendix A

Publications

We have published a number of works which have served as precursors to this work, and which have directly contributed to this work. In this section we list and briefly summarise each of these in chronological order.

1. **DCU at the TREC 2008 Blog Track**

   Adam Bermingham, Alan F. Smeaton, Jennifer Foster and Deirdre Hogan

   In: TREC 2008 - Text REtrieval Conference, Gaithersburg, MD.

   This notebook paper details our work for the opinion retrieval tasks at TREC and was our first foray into sentiment analysis. We used a hybrid approach of supervised classification using token-based features and syntactic features fused with scores from a sentiment lexicon. Our system was the third-best performing system for both the opinion-finding task and the polarity task.

2. **Combining Social Network Analysis and Sentiment Analysis to Explore the Potential for Online Radicalisation**

   Adam Bermingham, Maura Conway, Lisa McInerney, Neil O'Hare and Alan F. Smeaton

   In: ASONAM 2009 - Advances in Social Networks Analysis and Mining, 20-22 July, 2009, Athens, Greece.

   This research was a collaboration with the School of Law and Government in DCU

wherein we applied sentiment analysis and social network analysis to the problem of characterising dissident activity on social networks. We used an unsupervised sentiment classifier to identify sentiment and make observations particularly around the sentiment of different genders towards topics related to countries, organisations, figures and religions in the Middle East finding differences between genders in their sentiment towrds states and religions. This has inspired the demographic aspect to our thesis work.

3. **A study of Inter-annotator Agreement for Opinion Retrieval**
   Adam Bermingham and Alan F. Smeaton
   In: SIGIR 2009 - The 32nd Annual ACM SIGIR Conference, 20-22 July 2009, Boston, USA. ISBN 978-1-60558-483-6

   For this work we delved into the TREC relevance judgements for sentiment. We developed our own annotation methodology and tools, and ran a study with several sentiment annotators. We found a moderate level of agreement between annotators at a document level, yet wildly varied annotations for sentence-level annotations. We also found the mixed class to be highly prone to inconsistent interpretation.

4. **Topic-dependent Sentiment Analysis of Financial Blogs**
   Neil O'Hare, Michael Davy, Adam Bermingham, Paul Ferguson, Páraic Sheridan, Cathal Gurrin and Alan F. Smeaton,
   In: TSA 2009 - 1st International CIKM Workshop on Topic-Sentiment Analysis for Mass Opinion Measurement, 6 November 2009, Hong Kong, China. ISBN 978-1-60558-805-6

   In this paper we describe methodology for annotation of sentiment in financial blogs, an evolution from our work on annotating the content from the TREC blog corpus. Our experiments concern binary and ternary sentiment classification. One of the main contributions of this paper was demonstrating that considering topically relevant subsections significantly improves classifier performance.

5. **Exploring the Use of Paragraph-level Annotations for Sentiment Analy-**

sis of Financial Blogs

Paul Ferguson, Neil O'Hare, Michael Davy, Adam Bermingham, Páraic Sheridan, Cathal Gurrin and Alan F. Smeaton

In: WOMAS 2009 - Workshop on Opinion Mining and Sentiment Analysis, 13 November 2009, Seville, Spain.

In this publication we looked at annotating sentiment at a subdocument level, in this case for financial sentiment contained in blogs. We found that paragraph annotations provide better training data than full document texts but that this is outperformed by using relevance techniques to identify topically relevant subdocuments as training. Finding that classifying smaller topically relevant sections of text was easier contributed to us turning to short form domains as a potential avenue for sentiment research.

6. **Crowdsourced Real-world Sensing: Sentiment Analysis and the Real-time Web**

   Adam Bermingham and Alan F. Smeaton

   In: AICS 2010 - Sentiment Analysis Workshop at Artificial Intelligence and Cognitive Science, 30 August - 1 September 2010, Galway, Ireland.

   In this paper we review research in the area of real-time sentiment analysis. We provide motivation for progressing sentiment analysis in the context of the real-time web and describe our position on how the research field should progress.

7. **Classifying Sentiment in Microblogs: is Brevity an Advantage?**

   Adam Bermingham and Alan F. Smeaton

   CIKM 2010 - 19th International Conference on Information and Knowledge Management, 26-30 October 2010, Toronto, Canada. ISBN 978-1-4503-0099-5

   In this paper we described the experiments conatined in this thesis (Chapter 4) which address the problem of classifying sentiment in microblog posts. With a revised version of our annotation tool we conducted our third significant annotation

effort. We compared our classifcation results with those from other textual domains finding that brevity is mostly advantageous for sentiment classification and making other observations about the the classification such as the discriminability of various features.

8. **On Using Twitter to Monitor Political Sentiment and Predict Election Results**

   Adam Bermingham and Alan F. Smeaton

   In: SAAIP - Sentiment Analysis where AI meets Psychology workshop at the International Joint Conference on Natural Language Processing (IJCNLP) November 13, 2011, Shangri-La Hotel, Chiang Mai, Thailand. (in press)

   This work used the microblog data we collected, annotated and analysed for the Irish General Election (as described in Chapter 6) to test models for predicting the overall election result. We used a number of novel measures, inspired by previous research. We found that volume-based measures have a stronger predictive quality than sentiment, though both are helpful. We also present an overview of the GE11 Twitter Tracker.

Appendix B

Sentiment Annotation Guidelines

The following pages contain the guidelines that we distributed to annotators during the document labelling phase as described in Chapter 4. The versions of the guidelines here are the final versions used and had been refined during preliminary rounds of sample annotations and annotator training. These guidelines were also used during the training data creation phase for our real-time studies in Chapter 6, with additional topic-specific examples.

1st April 2010

Sentiment Annotation Guidelines

These guidelines contain important information on the definition of sentiment, description of the task and the annotation classes. All participants should have this document convenient when completing annotations.

Introduction

In sentiment annotation, we are looking for sentiment contained in documents towards various sentiment targets or topics. Each participant will be provided with a series of document-topic pairs – a sentiment target paired with a text document. For each pair, the participant should select one of the annotation classes.

Topics

Each topic is provided as a sentiment target (e.g. "Barack Obama"). This is followed by an annotation guide which identifies what should constitute **relevance** and **sentiment** for that topic.

In general we consider the topics to be a "sum of their parts". For example, this means that references to and sentiment towards a product also includes references to its features, its marketing and distribution but not to other products produced by the same company, employees of that company etc. However, if the target topic is a company, then relevant references to and sentiment towards the company, the company's employees as well as the company's products, the product's features etc. would be considered relevant.

This can get confusing, so each topic has four fields to aid the annotation process:

Topic Attribute	Description	Example
Topic	The target topic.	Bono
Relevance Guideline	A description of what constitutes relevance for this topic.	Documents which reference performer and political activist Bono are considered relevant.
Sentiment Guideline	A description of what constitutes relevant sentiment for this topic.	Relevant sentiment includes sentiment towards Bono, his music, his performances, his political acts or his personal life.

Description	A description of the topic from Wikipedia which you may optionally read if you are not familiar with the topic.	Paul David Hewson, KBE (born 10 May 1960), most commonly known by his stage name Bono, is an Irish singer and musician, best known for being the main vocalist of the Dublin-based rock band U2. Bono was born and raised in Dublin, Ireland, and attended Mount Temple

It is vital for this experiment that the anntotators annotate the topics exactly as specified in the topic description.

It is worth noting that the topics and documents date from February to May 2009.

Annotation Classes

The annotation classes are as follows:

Relevant & Neutral	Relevant to target topic but no sentiment towards topic expressed
Relevant & Positive	Predominantly positive sentiment towards topic expressed
Relevant & Negative	Predominantly negative sentiment towards topic expressed
Relevant & Mixed	Positive and negative sentiment towards topic expressed
Not-relevant	This document is not relevant to the topic
Unclear	There is not enough information to annotate or I am unsure how to annotate
Unannotatable	Spam, Inappropriate, Non-English, etc

Each document should be considered as a whole, rather than the sum of individual sentences or phrases. You may open URLs contained in the documents if you wish, though the annotation should remain based on the content of the original document. Note that a single document may be annotated differently for different topics.

You may not go back and correct an annotation. If you feel you have made an incorrect annotation, do not try to compensate or provide a further incorrect annotation in aid of consistency. Each document should be viewed as an

independent annotation so continue with the annotations annotating each as best as you can with respect to these guidelines.

What is Relevance?

A document is considered relevant if it references the target topic as described in the topic description and guidelines. If a document references the topic without explicitly discussing the topic or providing information about it, the document would be still be considered relevant.

Some special cases:

i) Non-English documents should always be marked "Unannotatable", even if it is clear that they reference the topic in another language.
ii) A topic mentioned in a URL is not considered relevant.
iii) Be careful to read the whole document. Often relevant information may be found in hashtags or short interjections.
iv) If an author is citing another source which references the topic, either by retweeting or direct quote, this is considered relevant.
v) Promotional tweets (ads) may constitute legitimate relevant documents and may bear sentiment. Documents which are obviously spam however should be marked "Unannotatable".

What is Sentiment?

Sentiment is defined as **positive** and **negative (i) opinions**, **(ii) emotions**, and **(iii) evaluations** towards given **topics.**

A document **is** considered to contain relevant sentiment if the document author expresses sentiment towards the topic and/or cites sentiment from another source towards the topic.

Although it is not necessary to annotate the type of sentiment, it may be helpful to keep the 3 types in mind when completing annotations.

(i) Opinion

*What **is** opinion sentiment?*
An opinion is an expression of the positive and/or negative subjective viewpoint of the author with respect to the target topic. For example: "I really like the new iPhone" would be positive towards the topic "iPhone".

*What **is not** opinion sentiment?*
Subjective views which are not polarised (positive or negative) are not considered to contain sentiment. For example: "I think the iPhone is a different colour to the last version" would be considered neutral towards the topic "iPhone".

(ii) Emotion

*What **is** emotion sentiment?*
Emotion is a positive or negative emotion expressed which has been induced by the topic. For example: “Obama's new healthcare bill makes me sad” would be negative towards the topic “Barrack Obama”.

*What **is not** emotion sentiment?*
An emotion which happens to co-occur with the topic does not necessarily constitute sentiment. For example: “In a good mood today. Heading to see 'The Hurt Locker' tonight” Would be considered neutral towards the topic “The Hurt Locker”.

(iii) Evaluation/Speculation

*What **is** evaluation sentiment?*
Evaluation is the assessment of a topic as positive or negative without necessarily qualifying it as your opinion. For example: “The Irish Economy is particularly weak at the moment”. A comment on an expectation or speculation is also considered sentiment. For example: “Looking forward to seeing Watchmen on Friday” (*expectation*), “The Irish Economy will improve next quarter” (*speculation*).

*What **is not** evaluation sentiment?*
Reporting of facts is not considered an evaluation unless it is being used to specifically illustrate an evaluation. For example: “Ireland are winning by a goal” would be considered neutral whereas “Ireland are playing well – they are up by a goal” would be considered positive.

A Note on Twitter

The documents presented in this experiment are microblog posts from the microblogging service “Twitter”[1]. Twitter is a service which allows users to publish short messages to those who “follow” them. If you are not familiar with Twitter, please take note of the following conventions.

Names

Users are referred to by prepending the character '@' to a username. For example: “In the Helix having a cup of coffee with @adambermingham”. If a user is referenced at the start of a message, this denotes that the message is specifically addressed to that user and may be part of a longer conversation thread.

Hashtags

1 http://www.twitter.com

Messages also sometimes contain the character '#' followed by a word. This is a convention for tagging the message as relevant to some entity, event, idea etc. For example: “This is the worst set of features for an Apple product yet! #iPhone”.

Retweet

If a user starts a message with “RT @<username>...” (sometimes “via @<username>”), this is a way of attributing a tweet to the original author and signifying that you are redistributing it among your followers.

Appendix C

Experiment Materials

C.1 Participant Instructions for Simulated Real-time Study

The following pages contain the instructions we provided to users for our simulated real-time scenarios in Chapter 5. The Channel S system prompted the users to answer general questions and to give feedback between topics. These survey questions were identical to those we use in our real-time studies (See Sections C.3 and C.4 for details).

Thesis Experiment
Participant Instructions

13th October 2010

Adam Bermingham

Introduction

Thanks for volunteering to participate in my thesis experiments. Please read the experiment details on this page carefully. The whole experiment should take less than an hour. If there are any problems, bugs or issues, please contact me on 0866067120 / abermingham@computing.dcu.ie / @adambermingham.

Experiment

The purpose of this task is to help develop a system for real-time search. In real-time search, the user enters a search query (or 'topic'), for example "mid-term election", and then waits for documents to appear relevant to that query. Unlike traditional web search, a topic may produce thousands of new relevant documents per minute, many more than a user can manage to read in real-time. A subset of these relevant documents must therefore be selected to show to the user. In this experiment, we simulate this scenario with 16 predefined queries. Your feedback will be used to help decide how best to reduce the number of documents down to a manageable amount.

For each topic you will first be presented with a description of that topic from Wikipedia. This is followed by a stream of documents relevant to that topic. Each document in the stream will appear at regular intervals, in chronological order, starting with the earliest. There will be 12 documents for each topic and after each topic, you will be asked to complete a short questionnaire. The topics will be preceded by one training topic for you to familiarise yourself with the task.

If you like a particular document in the stream and/or would like to see more of similar documents in your streams click the thumbs up icon beside the document. This is similar to 'liking' a link on facebook or a comment on blog. Likewise, if you dislike a document and/or think such documents are best avoided, click the thumbs down icon. Other documents you will want to neither dislike nor like; these do not require an action on your part. It should be noted that this type of user feedback is different from relevance feedback. Most of the documents in the streams will be relevant; however, we want to know which of these relevant documents you would prefer (and rather avoid) when you run a real-time query.

Twitter

The data used in this experiment is from the popular microblogging service, Twitter. Twitter allows users to publish short messages or 'tweets' online in real-time for others to see. The messages you see in the streams were authored by users in spring 2009. The topics also date from that period.

For those of you not familiar with Twitter, there are a couple of conventions to note. If a term is preceded with a hash character ('#'), this is considered to be a user tagging their post, e.g. #funnyjoke or #golf. If a term is preceded by an at character ('@') this is a Twitter username. If a username is the first term in a message, this signifies that the message is addressed to that user e.g. "@adambermingham really digging your neat experiment!".

Housekeeping

The task must be completed in one session. There is no support for the back button so please don't try to backtrack! All testing and development was done in Chrome but it should be fine in other browsers. Please send me an email when you have completed the experiment so I can check everything is in order.

C.2 Ethics Notification Form

The following pages contain the form for low-risk ethics notification we submitted to the university Research Ethics Committee in advance of carrying out our user studies. In this submission we also included some of the materials from the experiment, such as the informed consent form, plain language statement and questionnaires.

REC/2010/___

Research Ethics Committee: Notification Form for Low-Risk Projects and Undergraduate Dissertations

DCU Research Ethics Committee has introduced a procedure for notification to the committee of
1. low-risk social research projects, in which personal information that is deemed not sensitive is being collected by interview, questionnaire, or other means
2. dissertations on undergraduate programmes in all disciplines.

The committee requires researchers to concisely answer the following questions within this form (before the project starts):

Project Title:
Examining the Role of Sentiment in a Real-time Search Scenario

Applicant Name and E-mail:
Adam Bermingham, abermingham@computing.dcu.ie

If a student applicant, please provide the following:
Level of Study (Undergrad/Taught MSc/Research MSc/Phd): PhD
Supervisor Name and E-mail: Alan Smeaton, asmeaton@computing.dcu.ie

Questions:

1. Provide a lay description of the proposed research (approx. 300wds):

Much of today's Internet is taken up with content generated by the average user in the form of websites such as blogs, social networks and forums. By analysing what people write, we can understand how various entities of interest such as people, brands and products are perceived. This is an area of research called Sentiment Analysis – the automated analysis of opinions, emotions and evaluations from text.

Recently a user-generated content service which has seen huge popularity is the microblogging site, Twitter. Twitter allows users to publish and read short messages online. If a user wants to find out about an event for example, they may run a search on the Twitter website and the most recent Twitter posts related to that event will be retrieved. Due to the high volume of content on Twitter, there may be more than 10 new posts related to a particular event written every second.

A common practice is to monitor services such as Twitter while watching television. This allows viewers to observe an online social context while watching the television program. This is known as "second screen viewing". The problem is that with a lot of these events it is not clear how best to deliver the more appropriate messages to the user. After all, they couldn't possibly read all of the content being generated.

This research is proposing the use of Sentiment Analysis to filter the posts that a user observes in their stream. To that end, we are conducting a user study to get feedback data on how a variety of Sentiment-based algorithms perform in terms of delivering documents to users. These algorithms will test if users prefer to see subjective content or not or if they prefer negative or positive content, for example. The television programs we will use for our study are the penultimate and ultimate X-Factor live shows on the 4th and 11th of December.

2. Detail your proposed methodology (1 page max.):

The user study will take place on the 4th and 11th of December. On these days, 20 participants will be required to attend a room in the School of Computing in DCU to use the experiment system. In total 40 participants will participate over the two days.

Firstly the participants will be required to attend an orientation presentation where they will be introduced to Twitter, the X-Factor and the experiment system. Thereafter they will be given a short questionnaire to establish certain demographic details such as age, level of education, degree of familiarity with Twitter and the X-Factor. The will also be asked for their opinion towards a number of the sentiment targets in the X-Factor, ie, the acts, judges and the xfactor itself. Name will not be recorded and each user will be referred to internally with a unique ID.

The system itself is a web application. The system searches on Twitter for new messages related to the X-Factor. These are stored in a database. This database is periodically checked for new messages which mention any of the acts on the show. Any document which mentions an act is then classified for sentiment towards that act ie as neutral, positive, or negative. This requires a number of text processing steps, feature extraction and storing sentiment results. The sentiment classifiers are Multinomial Naive Bayes classifiers which will previously have been trained on Twitter data annotated for sentiment from previous X-Factor shows.

Each user will log into the web application as the X-Factor starts. Throughout the show they will be displayed a stream of Twitter messages. Beside each message is a thumbs up and a thumbs down icon – similar to the type an Internet user might see on a social network or in the comments in a blog. The users are told that they should thumbs up documents that they like ie would want to see more of in the stream, and thumbs down documents they dislike, ie types of documents they would like to see less of in their stream. At periodic intervals, the streams will pause and the users will be asked to fill out a short survey on the perceived quality of the most recent stream of tweets.

Meanwhile in the back end, the system will be choosing the tweets to give to each user in a systematic way. At each point in time, the users will be allocated one of a set of sentiment-based algorithms which will be used to determine which messages get shown to the user. The users will told that their feedback will be used to help design the real-time system but they will be unaware of the sentiment aspect of the experiment or the rotation of the algorithms. The algorithms are allocated to users in a latin squares arrangement. For example, for the period of time until the first survey, a user might be allocated the algorithm “Only Show Positive Messages”. After 10 minutes, they will then fill out a survey asking whether they found the preceding stream insightful, interesting, and informative and asking them to rate the overall quality of the stream. They then will resume annotating the stream but their documents might now be chosen using the algorithm “Show No Opinionated Messages”. Throughout the experiment they will rotate through the algorithms at these 10 minute intervals.

After the show(s) are complete, each user will be asked to fill out an exit survey asking general questions about the perceived performance of the system.

Our evaluation will use the survey data, coupled with the application feedback (thumbs up/thumbs down) data to ascertain the perceived performance of each of the set of sentiment algorithms compared to a control – a random sampling of the stream which will be used as one of the algorithms in rotation. Our hypothesis is the patterns will be evident in the user feedback and that by using Sentiment Analysis we can provide a better quality real-time stream to the user.

3. Detail the means by which potential participants will be recruited:

The participants will be found by emailing the following email lists and inviting participation - CLARITY email, CDVP email, School of Computing postgraduates, School of Electronic Engineering postgraduates.

4. How will the anonymity of the participants be respected? The user's name is not recorded and will not be used in any published data or papers.
5. What risks are researchers or participants being exposed to, if any? There are no risks to researchers or participants.
6. Have approval/s have been sought or secured from other sources? Yes/No If Yes, give details: No
7. Please confirm that the following forms are attached to this document: Informed Consent Form Yes/No Yes Plain Language Statement Yes/No Yes If not, explain why:

NB – The application should consist of <u>one file only</u>, which incorporates all supplementary documentation. The completed application must be proofread and spellchecked before submission to the REC. All sections of the form should be completed. Applications which do not adhere to these requirements will not be accepted for review and will be returned directly to the applicant.

The administrator to the Research Ethics Committee will assess, on receiving such notification, whether the information provided is adequate and whether any further action is necessary. Please complete this form and e-mail to fiona.brennan@dcu.ie

<u>Please note:</u> Project supervisors of dissertations on undergraduate programmes have the primary responsibility to ensure that students do not take on research that could expose them and the participants to significant risk, such as might arise, for example, in interviewing members of vulnerable groups such as young children.

In general, please refer to the Common Questions on Research Ethics Submissions for further guidance on what research procedures or circumstances might make ethical approval necessary (http://www.dcu.ie/internal/research/questions_ethics_submissions.pdf)

C.3 Participant Materials for Real-time Study - The X Factor

The following pages contain the booklet of materials that each participant received for the X Factor experiment (Chapter 6). Additional *Live Survey* pages were supplied, but are omitted here for brevity.

X-Factor Experiment

Introduction

Hi – thanks for volunteering to participate in the CLARITY X-Factor experiment. As well as using the Channel S system for monitoring X-Factor tweets during the show, this set of forms and surveys must be filled by the participants at various stages before, during and after the X-Factor:

Before the X-Factor starts:

- ☐ Read the Plain Language Statement
- ☐ Complete the Informed Consent Form
- ☐ Complete the Participant General Questionnaire
- ☐ Ensure you are able to log into the system and understand the task as explained

During the X-Factor:

- ☐ Complete a survey sheet entitled "Live Survey" each time you are prompted to by the system. Be sure to note the time on the form. You may not have to use all of the Live Surveys supplied.

After the X-Factor

- ☐ Complete the Closing Survey
- ☐ Return completed surveys, forms and questionnaires to the experiment coordinator.

If there are any problems, or you require any clarification, please ask a coordinator or contact Adam Bermingham on 086 606 71 20 or abermingham@computing.dcu.ie

Thank you again for participating!

DUBLIN CITY UNIVERSITY

Plain Language Statement

Real-time Search User Evalutation for a Live Television Event
CLARITY: Centre for Sensor Web Technology,
Centre for Digital Video Processing (CDVP)

Principal Investigator: Prof. Alan Smeaton
asmeaton@computing.dcu.ie
700 5262

Investigator: Mr. Adam Bermingham
abermingham@computing.dcu.ie
700 6840

The study will consist of an orientation and training session (approximately 1hr) followed be the experiment itself (approximately 2 hrs) totaling approximately 3hrs. The study will take place twice, on the 11^{h} of December and the 12th December. You are required to attend only one of these days, as agreed during the volunteering process.

You will be introduced to Twitter, the X-Factor and the experiment system. You will be asked to fill out a short demographic survey. Your details will be anonymised. You will use the experiment system for the duration of the X-Factor live show, providing feedback to the system as instructed in the training session. You will also fill out periodic questionnaires at (approximately) 10 minute intervals regarding the performance of the system. Finally, you will be asked to fill out introductory and exit surveys concerning overall performance of the system and concerning your previous experience of technologies and your knowledge and opinion of various aspects of the X-Factor. The experiment will be photographed and videotaped.

There are no risks from participating in this study.

There are no direct benefits promised to you apart from the opportunity to use the real-time system and learn about the associated technologies.

Your name will not be stored with your demographic and survey information. You will be assigned an ID number which will be used in all references to your data which will be stored in a secure database in DCU. Your identity will not be revealed or published.

Your data, including photographs and video will be kept for a maximum of 5 years after which it will be shredded. Photographs or videos of you will not be published or made publicly available.

There are no risks to me from participating in this study. Participation is voluntary and there is no penalty for leaving the study prematurely.

If participants have concerns about this study and wish to contact an independent person,

please contact:

The Secretary, Dublin City University Research Ethics Committee, c/o Office of the Vice-President for Research, Dublin City University, Dublin 9. Tel 01-7008000

DUBLIN CITY UNIVERSITY

Informed Consent

Real-time Search User Evalutation for a Live Television Event
CLARITY: Centre for Sensor Web Technology,
Dublin City University

Principal Investigator: Prof. Alan Smeaton
asmeaton@computing.dcu.ie
700 5262

Investigator: Mr. Adam Bermingham
abermingham@computing.dcu.ie
700 6840

The purpose of this research is to gather feedback from users as they use a real-time search system while simultaneously watching a live television event.

I will attend an orientation and training session (approximately 1hr) followed be the experiment itself (approximately 2 hrs) totaling approximately 3hrs. I will attend these events on either the 11th of December or the 12th December as was agreed in the volunteering process.

I will be introduced to Twitter, the X-Factor and the experiment system. I will fill out a demographic survey. Any details I give will be anonymised and will not be stored with my name. I will use the experiment system for the duration of the X-Factor live show, providing feedback to the system as instructed in the training session. I will also fill out periodic surveys at approximately 10 minute intervals regarding the performance of the system. I will also fill out introductory and exit surveys concerning overall performance of the system and concerning my previous experience of technologies and my knowledge and opinion of various aspects of the X-Factor.

Participant – please complete the following (Circle Yes or No for each question)

Have you read or had read to you the Plain Language Statement	*Yes/No*
Do you understand the information provided?	*Yes/No*
Have you had an opportunity to ask questions and discuss this study?	*Yes/No*
Have you received satisfactory answers to all your questions?	*Yes/No*
Are you aware that the study may be photographed and videotaped?	*Yes/No*

There are no risks to me from participating in this study. Participation is voluntary and there is no penalty for leaving the study prematurely.

My name will not be stored with my demographic and survey information. I will be assigned an ID number which will be used in all references to my data which will be stored in a secure database in DCU. My identity will not be revealed or published.

I have read and understood the information in this form. My questions and concerns have been answered by the researchers, and I have a copy of this consent form. Therefore, I consent to take part in this research project

Participants Signature: _____

Name in Block Capitals: _____

Witness: _____

Date: _____

X-Factor Experiment
General Questionnaire

Participant Profile

Name:

ID:

What is your **education** level?

- ☐ Undergraduate student / no degree
- ☐ Postgraduate student / you have a primary degree
- ☐ Researcher / you have an advanced degree
- ☐ Faculty or research staff

What is your **gender**?

- ☐ Male
- ☐ Female

What is your **age**?

- ☐ 18-24
- ☐ 25-34
- ☐ 35-44
- ☐ 45-54
- ☐ 65+

Twitter Usage

How often do you **post** messages on Twitter?

- ☐ More than once a day
- ☐ Once a day
- ☐ A few times a week
- ☐ Once a week
- ☐ Never

How often do you perform **searches** using Twitter?

- ☐ More than once a day
- ☐ Once a day
- ☐ A few times a week
- ☐ Once a week
- ☐ Never

How often do you **read** messages on Twitter?

- ☐ More than once a day
- ☐ Once a day
- ☐ A few times a week
- ☐ Once a week
- ☐ Never

Overall how would you rate your **familiarity** with Twitter?

- ☐ not at all familiar
- ☐ Slightly familiar
- ☐ Somewhat familiar
- ☐ Moderately familiar
- ☐ Extremely familiar

X-Factor

How **frequently** do you watch the X-Factor?

- ☐ Always
- ☐ Very Often
- ☐ Sometimes
- ☐ Rarely
- ☐ Never

Which of the following best describes your opinion towards the **X-Factor**?

- ☐ I have mostly positive opinions towards the X-Factor
- ☐ I have mostly negative opinions towards the X-Factor
- ☐ I am familiar with the X-Factor but have neither positive nor negative opinions towards the X-Factor
- ☐ I am unfamiliar with the X-Factor

Which of the following best describes your opinion towards the contestant, **Matt Cardle**?

- ☐ I have mostly positive opinions towards Matt Cardle
- ☐ I have mostly negative opinions towards Matt Cardle
- ☐ I am familiar with Matt Cardle but have neither positive nor negative opinions towards Matt Cardle
- ☐ I am unfamiliar with Matt Cardle

Which of the following best describes your opinion towards the contestant, **Rebecca Ferguson**?

- ☐ I have mostly positive opinions towards Rebecca Ferguson
- ☐ I have mostly negative opinions towards Rebecca Ferguson
- ☐ I am familiar with Rebecca Ferguson but have neither positive nor negative opinions towards Rebecca Ferguson
- ☐ I am unfamiliar with Rebecca Ferguson

Which of the following best describes your opinion towards the contestant, **One Direction**?

- ☐ I have mostly positive opinions One Direction
- ☐ I have mostly negative opinions One Direction
- ☐ I am familiar with One Direction but have neither positive nor negative opinions towards One Direction
- ☐ I am unfamiliar with One Direction

Which of the following best describes your opinion towards the contestant, **Cher Lloyd**?

- ☐ I have mostly positive opinions towards Cher Lloyd
- ☐ I have mostly negative opinions towards Cher Lloyd
- ☐ I am familiar with Cher Lloyd but have neither positive nor negative opinions towards Cher Lloyd
- ☐ I am unfamiliar with Cher Lloyd

Which of the following is your **favourite** act?

- ☐ Rebecca Ferguson
- ☐ One Direction
- ☐ Cher Lloyd
- ☐ Matt Cardle
- ☐ No opinion

Who do you **predict will win** the X-Factor overall?

- ☐ Rebecca Ferguson
- ☐ One Direction
- ☐ Cher Lloyd
- ☐ Matt Cardle
- ☐ No opinion

X-Factor Experiment
Live Survey

Time: [ ] Live Survey Number: [ ]

Please answer the following questions with respect to the preceding stream of tweets, i.e. the tweets you have seen since the last time you completed a live survey or if this is your first survey since the start of the show.

To what extent do you agree with each of the following statements?

	Strongly Disagree	Disagree	Neither Disagree nor Agree	Agree	Strongly Agree
The tweets in the preceding stream were **insightful**.	☐	☐	☐	☐	☐
The tweets in the preceding stream were **interesting**.	☐	☐	☐	☐	☐
The tweets in the preceding stream were **informative**.	☐	☐	☐	☐	☐

Overall how would you rate the content in the preceding stream?

Poor Excellent

☐ ☐ ☐ ☐ ☐ ☐ ☐

X-Factor Experiment

Closing Survey

Please answer the following questions with respect to all of the tweets you have seen during the show.

To what extent do you agree with each of the following statements?

	Strongly Disagree	Disagree	Neither Disagree nor Agree	Agree	Strongly Agree
The tweets in general were **insightful**.	☐	☐	☐	☐	☐
The tweets in general were **interesting**.	☐	☐	☐	☐	☐
The tweets in general were **informative**.	☐	☐	☐	☐	☐

Overall how would you rate the content in general?

Poor Excellent

☐ ☐ ☐ ☐ ☐ ☐ ☐

You have now completed the experiment – please return your surveys and questionnaires to the event coordinator. Thank you for your time.

C.4 Participant Materials for Real-time Study - The Leaders' Debate

The following pages contain the booklet of materials that each participant received for the Leaders' Debate experiment (Chapter 6). Additional *Live Survey* pages were supplied, but are omitted here for brevity.

GE11 Experiment
Introduction

Hi – thanks for volunteering to participate in the CLARITY Leader's Debate experiment. As well as using the Channel S system for monitoring tweets during the debate, this set of forms and surveys must be filled by the participants at various stages before, during and after the debate:

Before the debate starts:

- ☐ Read the Plain Language Statement
- ☐ Complete the Informed Consent Form
- ☐ Complete the Participant General Questionnaire
- ☐ Ensure you are able to log into the system and understand the task as explained

During the debate:

- ☐ Complete a survey sheet entitled "Live Survey" each time you are prompted to by the system. Be sure to note the time on the form. You may not have to use all of the Live Surveys supplied.

After the debate

- ☐ Complete the Closing Survey (found on the last page here)
- ☐ Return completed surveys, forms and questionnaires to the experiment coordinator.

If there are any problems, or you require any clarification, please ask a coordinator or contact Adam Bermingham on 086 606 71 20 or abermingham@computing.dcu.ie

Thank you again for participating!

DUBLIN CITY UNIVERSITY

Plain Language Statement

Real-time Search User Evalutation for a Live Television Event
CLARITY: Centre for Sensor Web Technology,
Centre for Digital Video Processing (CDVP)

Principal Investigator: Prof. Alan Smeaton
asmeaton@computing.dcu.ie
700 5262

Investigator: Mr. Adam Bermingham
abermingham@computing.dcu.ie
700 6840

The study will consist of an orientation and training session (approximately 1hr) followed be the experiment itself (approximately 1hr) totaling approximately 2hrs. The study will take place on the 14^{h} of February. You are required to attend on this date, as agreed during the volunteering process.

You will be introduced to Twitter, the Leaders' Debate and the experiment system. You will be asked to fill out a short demographic survey. Your details will be anonymised. You will use the experiment system for the duration of the debate broadcast, providing feedback to the system as instructed in the training session. You will also fill out periodic questionnaires at (approximately) 15 minute intervals regarding the performance of the system. Finally, you will be asked to fill out introductory and exit surveys concerning overall performance of the system and concerning your previous experience of technologies and your knowledge and opinion of various aspects of the the debate. The experiment will be photographed and videotaped.

There are no risks from participating in this study.

There are no direct benefits promised to you apart from the opportunity to use the real-time system and learn about the associated technologies.

Your name will not be stored with your demographic and survey information. You will be assigned an ID number which will be used in all references to your data which will be stored in a secure database in DCU. Your identity will not be revealed or published.

Your data, including photographs and video will be kept for a maximum of 5 years after which it will be shredded. Photographs or videos of you will not be published or made publicly available.

There are no risks to me from participating in this study. Participation is voluntary and there is no penalty for leaving the study prematurely.

If participants have concerns about this study and wish to contact an independent person,

please contact:

The Secretary, Dublin City University Research Ethics Committee, c/o Office of the Vice-President for Research, Dublin City University, Dublin 9. Tel 01-7008000

DUBLIN CITY UNIVERSITY

Informed Consent

Real-time Search User Evalutation for a Live Television Event
CLARITY: Centre for Sensor Web Technology,
Dublin City University

Principal Investigator: Prof. Alan Smeaton
asmeaton@computing.dcu.ie
700 5262

Investigator: Mr. Adam Bermingham
abermingham@computing.dcu.ie
700 6840

The purpose of this research is to gather feedback from users as they use a real-time search system while simultaneously watching a live television event.

I will attend an orientation and training session (approximately 1hr) followed be the experiment itself (approximately 1 hrs) totaling approximately 2hrs. I will attend this event on the 14th February as was agreed in the volunteering process.

I will be introduced to Twitter, the Leaders' Debate and the experiment system. I will fill out a demographic survey. Any details I give will be anonymised and will not be stored with my name. I will use the experiment system for the duration of the debate broadcast, providing feedback to the system as instructed in the training session. I will also fill out periodic surveys at approximately 15 minute intervals regarding the performance of the system. I will also fill out introductory and exit surveys concerning overall performance of the system and concerning my previous experience of technologies and my knowledge and opinion of various aspects of the debate.

Participant – please complete the following (Circle Yes or No for each question)

Have you read or had read to you the Plain Language Statement	*Yes/No*
Do you understand the information provided?	*Yes/No*
Have you had an opportunity to ask questions and discuss this study?	*Yes/No*
Have you received satisfactory answers to all your questions?	*Yes/No*
Are you aware that the study may be photographed and videotaped?	*Yes/No*

There are no risks to me from participating in this study. Participation is voluntary and there is no penalty for leaving the study prematurely.

My name will not be stored with my demographic and survey information. I will be assigned an ID number which will be used in all references to my data which will be stored in a secure database in DCU. My identity will not be revealed or published.

I have read and understood the information in this form. My questions and concerns have been answered by the researchers, and I have a copy of this consent form. Therefore, I consent to take part in this research project

Participants Signature: _____

Name in Block Capitals: _____

Witness: _____

Date: _____

GE11 Experiment
General Questionnaire

Participant Profile

Name:

ID:

What is your **education** level?

- ☐ Undergraduate student / no degree
- ☐ Postgraduate student / you have a primary degree
- ☐ Researcher / you have an advanced degree
- ☐ Faculty or research staff

What is your **gender**?

- ☐ Male
- ☐ Female

What is your **age**?

- ☐ 18-24
- ☐ 25-34
- ☐ 35-44
- ☐ 45-54
- ☐ 65+

Twitter Usage

How often do you **post** messages on Twitter?

- ☐ More than once a day
- ☐ Once a day
- ☐ A few times a week
- ☐ Once a week
- ☐ Never

How often do you perform **searches** using Twitter?

- ☐ More than once a day
- ☐ Once a day
- ☐ A few times a week
- ☐ Once a week
- ☐ Never

How often do you **read** messages on Twitter?

- ☐ More than once a day
- ☐ Once a day
- ☐ A few times a week
- ☐ Once a week
- ☐ Never

Overall how would you rate your **familiarity** with Twitter?

- ☐ not at all familiar
- ☐ Slightly familiar
- ☐ Somewhat familiar
- ☐ Moderately familiar
- ☐ Extremely familiar

GE11

How **frequently** do you watch current affairs television?

- ☐ Always
- ☐ Very Often
- ☐ Sometimes
- ☐ Rarely
- ☐ Never

Which of the following best describes your opinion towards the **General Election**?

- ☐ I have mostly positive opinions towards the General Election
- ☐ I have mostly negative opinions towards the General Election
- ☐ I am familiar with the General Election but have neither positive nor negative opinions towards the General Election
- ☐ I am unfamiliar with the General Election

Which of the following best describes your opinion towards tonight's **Leaders' Debate**?

- ☐ I have mostly positive opinions towards the Leaders' Debate
- ☐ I have mostly negative opinions towards the Leaders' Debate
- ☐ I am familiar with the Leaders' Debate but have neither positive nor negative opinions towards the Leaders' Debate
- ☐ I am unfamiliar with Leaders' Debate

Which of the following best describes your opinion towards Fianna Fáil leader, **Micheál Martin**?

- ☐ I have mostly positive opinions towards Micheál Martin
- ☐ I have mostly negative opinions towards Micheál Martin
- ☐ I am familiar with Micheál Martin but have neither positive nor negative opinions towards Micheál Martin
- ☐ I am unfamiliar with Micheál Martin

Which of the following best describes your opinion towards the Green Party leader, **John Gormley**?

- ☐ I have mostly positive opinions towards John Gormley
- ☐ I have mostly negative opinions towards John Gormley
- ☐ I am familiar with John Gormley but have neither positive nor negative opinions towards John Gormley
- ☐ I am unfamiliar with John Gormley

Which of the following best describes your opinion towards the Sinn Féin leader, **Gerry Adams**?

- ☐ I have mostly positive opinions towards Gerry Adams
- ☐ I have mostly negative opinions towards Gerry Adams
- ☐ I am familiar with Gerry Adams but have neither positive nor negative opinions towards Gerry Adams
- ☐ I am unfamiliar with Gerry Adams

Which of the following best describes your opinion towards the Fine Gael leader, **Enda Kenny**?

- ☐ I have mostly positive opinions towards Enda Kenny
- ☐ I have mostly negative opinions towards Enda Kenny
- ☐ I am familiar with Enda Kenny but have neither positive nor negative opinions towards Enda Kenny
- ☐ I am unfamiliar with Enda Kenny

Which of the following best describes your opinion towards the Labour leader, **Eamon Gilmore**?

- ☐ I have mostly positive opinions towards Eamon Gilmore
- ☐ I have mostly negative opinions towards Eamon Gilmore
- ☐ I am familiar with Eamon Gilmore but have neither positive nor negative opinions towards Eamon Gilmore
- ☐ I am unfamiliar with Eamon Gilmore

Which of the following is your **favourite** leader?

- ☐ John Gormley
- ☐ Micheál Martin
- ☐ Gerry Adams
- ☐ Enda Kenny
- ☐ Eamon Gilmore

Who do you **predict will be perceived the winner** of the debate?

- ☐ John Gormley
- ☐ Micheál Martin
- ☐ Gerry Adams
- ☐ Enda Kenny
- ☐ Eamon Gilmore

GE11 Experiment

Live Survey

Time: ☐ Live Survey Number: ☐

Please answer the following questions with respect to the preceding stream of tweets, i.e. the tweets you have seen since the last time you completed a live survey or if this is your first survey since the start of the show.

To what extent do you agree with each of the following statements?

	Strongly Disagree	Disagree	Neither Disagree nor Agree	Agree	Strongly Agree
The tweets in the preceding stream were **insightful**.	☐	☐	☐	☐	☐
The tweets in the preceding stream were **interesting**.	☐	☐	☐	☐	☐
The tweets in the preceding stream were **informative**.	☐	☐	☐	☐	☐

Overall how would you rate the content in the preceding stream?

Poor						Excellent
☐	☐	☐	☐	☐	☐	☐

GE11 Experiment

Closing Survey

Please answer the following questions with respect to all of the tweets you have seen during the show.

To what extent do you agree with each of the following statements?

	Strongly Disagree	Disagree	Neither Disagree nor Agree	Agree	Strongly Agree
The tweets in general were **insightful**.	☐	☐	☐	☐	☐
The tweets in general were **interesting**.	☐	☐	☐	☐	☐
The tweets in general were **informative**.	☐	☐	☐	☐	☐

Overall how would you rate the content in general?

Poor						Excellent
☐	☐	☐	☐	☐	☐	☐

You have now completed the experiment – please return your surveys and questionnaires to the event coordinator. Thank you for your time.

Appendix D

Topics

Here we present the topics we used in our experiments. Table D.1 contains the topics used in Chapter 4. Those topics we selected to use in Chapter 5 are **bolded**. The topics for our real-time user studies are contained in Tables D.2 and D.3. In both, further similar topics were used during training, though they are omitted here for brevity. We also omit the background descriptions which accompanied the topics during our experiments.

D.1 Topics for Sentiment Analysis Evaluation and Simulated Real-time User Studies

1	**Susan Boyle**	Documents which reference reality TV singer Susan Boyle are considered relevant.	Relevant sentiment includes sentiment towards Susan Boyle, her music, her performances or her personal life.	Entertainment
2	**Twilight**	Documents which reference any of the films in the Twilight Saga film series are considered relevant.	Relevant sentiment includes sentiment towards the Twilight films, their production or cast performances, or any associated products.	Entertainment
3	Leno	Documents which reference American talk show host Jay Leno are considered relevant.	Relevant sentiment includes sentiment towards Jay Leno, his shows, his performances or his personal life.	Entertainment

4	**Bono**	Documents which reference performer and political activist Bono are considered relevant.	Relevant sentiment sentiment includes sentiment towards Bono, his music, his performances, his political acts or his personal life.	Entertainment
5	Adam Lambert	Documents which reference reality TV performer Adam Lambert are considered relevant.	Relevant sentiment sentiment includes sentiment towards Adam Lambert, his music, his performances or his personal life.	Entertainment
6	**Watchmen**	Documents which reference the film Watchmen are considered relevant.	Relevant sentiment includes sentiment towards the film, its production or cast performances in the film or any associated products.	Entertainment
7	Rihanna	Documents which reference R&B performer Rihanna are considered relevant.	Relevant sentiment includes sentiment towards Rihanna, her music, her performances or her personal life.	Entertainment
8	Fargo	Documents which reference the film Fargo are considered relevant.	Relevant sentiment includes sentiment towards the film, its production or cast performances in the film or any associated products.	Entertainment
9	**Red Dwarf**	Documents which reference the science-fiction TV series Red Dwarf are considered relevant.	Relevant sentiment includes sentiment towards the series its production or cast performances in the series or any associated products.	Entertainment
10	Coachella	Documents which reference the music festival Coachella are considered relevant.	Relevant sentiment includes sentiment towards the festival, performances at the festival or services at the festival.	Entertainment
11	**man utd**	Documents which reference Manchester United Football Club are considered relevant.	Relevant sentiment includes sentiment towards the club, its management, owners, staff or team members is considered relevant sentiment.	Sports
12	**Celtics**	Documents which reference The Boston Celtics basketball team are considered relevant.	Relevant sentiment includes sentiment towards the team, its management, owners, staff or team members is considered relevant sentiment.	Sports

13	**Arsenal**	Documents which reference Arsenal Football Club are considered relevant.	Relevant sentiment includes sentiment towards the club, its management, owners, staff or team members is considered relevant sentiment.	Sports
14	Tiger Woods	Documents which reference golfer Tiger Woods are considered relevant	Relevant sentiment includes sentiment towards Tiger Woods, his career, performance or his personal life	Sports
15	**Lance Armstrong**	Documents which reference cyclist Lance Armstrong are considered relevant	Relevant sentiment includes sentiment towards Lance Armstrong, his cycling career, performance, his charity work or his personal life.	Sports
16	Curt Schilling	Documents which reference baseball player Curt Schilling are considered relevant.	Relevant sentiment includes sentiment towards Curt Schilling, his baseball career and performance or his personal life. Sentiment towards the Red Sox is not considered relevant sentiment.	Sports
17	**Mets**	Documents which reference baseball team The New York Mets are considered relevant.	Relevant sentiment includes sentiment towards the team, its management, owners, staff or team members is considered relevant sentiment.	Sports
18	Buffalo Bills	Documents which reference American football team The Buffalo Bills are considered relevant.	Relevant sentiment includes sentiment towards the team, its management, owners, staff or team members is considered relevant sentiment.	Sports
19	Terrell Owens	Documents which reference American football player Terrell Owens are considered relevant.	Relevant sentiment includes sentiment towards Terrell Owens, his career, performance or his personal life. Sentiment towards a team he has played for is not considered relevant sentiment.	Sports
20	Wales	Documents which reference the Welsh rugby union team are considered relevant.	Relevant sentiment includes sentiment towards the team, its management, staff or team members is considered relevant sentiment.	Sports

21	North Korea	Documents which reference the country North Korea are considered relevant.	Relevant sentiment includes sentiment towards its leadership, political policy or actions, culture or economy.	Politics & Government
22	NATO	Documents which reference the organisation NATO are considered relevant.	Relevant sentiment includes sentiment towards the NATO alliance and its members.	Politics & Government
23	Afghanistan War	Documents which reference the ongoing conflict in Afghanistan are considered relevant.	Relevant sentiment includes sentiment towards the current war in Afghanistan. Sentiment towards either side in the war is not considered relevant sentiment.	Politics & Government
24	Dave Ramsey	Documents which reference the politician Dave Ramsey are considered relevant.	Relevant sentiment includes sentiment towards Dave Ramsay, his policies, his actions or his personal life.	Politics & Government
25	Rush Limbaugh	Documents which reference radio host and political commentator Rush Limbaugh are considered relevant.	Relevant sentiment includes sentiment towards Rush Limbaugh, his commentary, his show or his personal life.	Politics & Government
26	Navy SEALS	Documents which reference the US Navy Seals are considered relevant.	Relevant sentiment includes sentiment towards the Navy Seals or members of the Navy Seals.	Politics & Government
27	Gordon Brown	Documents which reference UK Prime Minister Gordon Brown are considered relevant.	Relevant sentiment includes sentiment towards Gordon Brown, his actions, his policies or his personal life.	Politics & Government
28	Sanjay Gupta	Documents which reference neurosurgeon and former Surgeon General candidate Sanjay Gupta are considered relevant	Relevant sentiment includes sentiment towards Sanjay Gupta an dhis decision to opt out of the running for Surgeon General.	Politics & Government
29	**Obama**	Documents which reference American President Barack Obama are considered relevant.	Relevant sentiment includes sentiment towards Barack Obama, his actions, his policies or his personal life.	Politics & Government
30	budget	Documents which reference the United States budget are considered relevant.	Relevant sentiment includes sentiment towards the US budget, its delivery, its contents and its repercussions.	Politics & Government

31	Kindle	Documents which reference the Amazon e-reader the Kindle (any version) are considered relevant.	Relevant sentiment includes sentiment towards the Kindle, its features or any associated content or services. Sentiment towards Amazon in general is not considered relevant sentiment.	Products & Services
32	Wolfram Alpha	Documents which reference the knowledge engine Wolfram Alpha are considered relevant.	Relevant sentiment includes sentiment towards Wolfram Alpha, the service it provides or its features. Sentiment towards Stephen Wolfram or the company Wolfram Research is not considered relevant sentiment.	Products & Services
33	Guinness	Documents which reference the alcoholic beverage Guinness are considered relevant.	Relevant sentiment includes sentiment towards the beverage Guinness and experiences of consuming it. Sentiment towards Diageo is not considered relevant sentiment.	Products & Services
34	Pirate Bay	Documents which reference bittorrent website Pirate Bay are considered relevant.	Relevant sentiment includes sentiment towards Pirate Bay, its services or features. Sentiment towards Pirate Bay's staff is not considered relevant sentiment.	Products & Services
35	Skype	Documents which reference VoIP and IM software and service Skype are considered relevant.	Relevant sentiment includes sentiment towards Skype products and services are considered relevant. Sentiment towards the Skype Group or eBay is not considered relevant sentiment.	Products & Services
36	Sky News	Documents which reference the news content broadcaster Sky News are considered relevant.	Relevant sentiment includes sentiment towards Sky News, it's presenters, programmes, content and production. Sentiment towards BSkyB or other Sky channels or services is not considered relevant sentiment.	Products & Services

37	Nikon D5000	Documents which reference the camera Nikon D5000 are considered relevant.	Relevant sentiment includes sentiment towards the Nikon D5000 camera or its features. Sentiment towards the Nikon Corporation or other Nikon products or services is not considered relevant sentiment.	Products & Services
38	**Safari 4**	Documents which reference the Apple browser Safari version 4 are considered relevant.	Relevant sentiment includes sentiment towards Safari 4 or its features. Sentiment towards other version of the browser, other Apple products or Apple Inc. is not considered relevant sentiment.	Products & Services
39	**iPhone**	Documents which reference the Apple smartphone the iPhone are considered relevant.	Relevant sentiment includes sentiment towards iPhone or its features. Sentiment towards other Apple products or Apple Inc. is not considered relevant sentiment.	Products & Services
40	Spotify	Documents which reference the streaming music service and application Spotify are considered relevant.	Relevant sentiment includes sentiment towards the Spotify application, service or its features. Sentiment towards the company, Spotify itself is not considered relevant sentiment.	Products & Services
41	AIG	Documents which reference insurance company American International Group (AIG) are considered relevant.	Relevant sentiment includes sentiment towards AIG, its staff, services, products, stock or business.	Companies
42	Oracle	Documents which reference the IT company Oracle or their products or services are considered relevant.	Relevant sentiment includes sentiment towards Oracle, its staff, services, products, stock or business.	Companies
43	**Wal-Mart**	Documents which reference the retail corporation and chain Wal-Mart (or Walmart) are considered relevant.	Relevant sentiment includes sentiment towards Wal-Mart, its staff, retail shops, services, stock or business.	Companies
44	Sun Microsystems	Documents which reference IT company Sun Microsystems or its products and services are considered relevant.	Relevant sentiment includes sentiment towards Sun Microsystems, its staff, products, services, stock or business.	Companies

45	CNBC	Documents which reference broadcaster and content producer CNBC are relevant.	Relevant sentiment includes sentiment towards CNBC, its staff, presenters, production, content, programmes, stock or business.	Companies
46	Chrysler	Documents which reference car manufacturer Chrysler or cars they produce are considered relevant.	Relevant sentiment includes sentiment towards Chrysler, its staff, products, stock or business.	Companies
47	Lloyds	Documents which reference financial instiution Lloyds Bank or its services are considered relevant.	Relevant sentiment includes sentiment towards Lloyds, its staff, services, products, stock and its business.	Companies
48	**IBM**	Documents which reference IT corporation IBM or any of its products or services are considered relevant.	Relevant sentiment includes sentiment towards IBM, its staff, its products, services, its stock or its business.	Companies
49	Toyota	Documents which reference car manufacturer Toyota or the cars they produce are considered relevant.	Relevant sentiment includes sentiment towards Toyota, its staff, products, stock or business.	Companies
50	ACMA	Documents which reference broadcasting regulator ACMA are considered relevant.	Relevant sentiment includes sentiment towards ACMA, its policies, actions, staff or services.	Companies

Table D.1: Topics for our sentiment analysis evaluation in Chapter 4.

D.2 Topics for Real-time User Studies

2	Matt Cardle	Relevant documents are those which mention this act, or their performances.	Relevant sentiment is sentiment which is directed either towards the act ifself, their performance or an evaluation of their prospects in the competition.	act
3	Rebecca Ferguson	Relevant documents are those which mention this act, or their performances.	Relevant sentiment is sentiment which is directed either towards the act ifself, their performance or an evaluation of their prospects in the competition.	act
5	Cher Lloyd	Relevant documents are those which mention this act, or their performances.	Relevant sentiment is sentiment which is directed either towards the act ifself, their performance or an evaluation of their prospects in the competition.	act
6	One Direction	Relevant documents are those which mention this act, or their performances.	Relevant sentiment is sentiment which is directed either towards the act ifself, their performance or an evaluation of their prospects in the competition.	act
10	Dannii Minogue	Relevant documetns are those which mention this judge.	Relevant sentiment is sentiment which is directed either towards the judge themselves, their performance as mentor or statements they make.	judge
11	Simon Cowell	Relevant documetns are those which mention this judge.	Relevant sentiment is sentiment which is directed either towards the judge themselves, their performance as mentor or statements they make.	judge
12	Cheryl Cole	Relevant documetns are those which mention this judge.	Relevant sentiment is sentiment which is directed either towards the judge themselves, their performance as mentor or statements they make.	judge
13	Louis Walsh	Relevant documetns are those which mention this judge.	Relevant sentiment is sentiment which is directed either towards the judge themselves, their performance as mentor or statements they make.	judge

Table D.2: Topics for real-time X Factor study in Chapter 6.

0	Greens	All tweets which mention this party, are considered relevant.	Relevant sentiment is sentiment towards this party, its members, its policies or its prospects in the election.	party
1	Fianna Fáil	All tweets which mention this party, are considered relevant.	Relevant sentiment is sentiment towards this party, its members, its policies or its prospects in the election.	party
2	Fine Gael	All tweets which mention this party, are considered relevant.	Relevant sentiment is sentiment towards this party, its members, its policies or its prospects in the election.	party
3	Sinn Féin	All tweets which mention this party, are considered relevant.	Relevant sentiment is sentiment towards this party, its members, its policies or its prospects in the election.	party
4	Labour	All tweets which mention this party, are considered relevant.	Relevant sentiment is sentiment towards this party, its members, its policies or its prospects in the election.	party
75348	Eamon Gilmore	All tweets which mention this candidate are considered relevant.	Relevant sentiment is sentiment towards this candidate, their actions, their policies or their prospects in the election.	candidate
75353	Enda Kenny	All tweets which mention this candidate are considered relevant.	Relevant sentiment is sentiment towards this candidate, their actions, their policies or their prospects in the election.	candidate
75371	Gerry Adams	All tweets which mention this candidate are considered relevant.	Relevant sentiment is sentiment towards this candidate, their actions, their policies or their prospects in the election.	candidate
75424	John Gormley	All tweets which mention this candidate are considered relevant.	Relevant sentiment is sentiment towards this candidate, their actions, their policies or their prospects in the election.	candidate
77016	Micheál Martin	All tweets which mention this candidate are considered relevant.	Relevant sentiment is sentiment towards this candidate, their actions, their policies or their prospects in the election.	candidate

Table D.3: Topics for real-time Leaders' Debate study in Chapter 6.

Bibliography

Adomavicius, G. and Tuzhilin, A. (2005). Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE transactions on knowledge and data engineering*, pages 734–749.

af Segerstad, Y. H. (2003). *Use and Adaptation of Written Language to the Conditions of Computer-Mediated Communication*. PhD thesis, Goteborg University, Sweden.

Agarwal, A., Xie, B., Vovsha, I., Rambow, O., and Passonneau, R. (2011). Sentiment analysis of twitter data. In *Proceedings of the Workshop on Languages in Social Media*, pages 30–38. Association for Computational Linguistics.

Agarwal, S., Godbole, S., Punjani, D., and Roy, S. (2007). How much noise is too much: A study in automatic text classification. In *Industrial Conference on Data Mining (ICDM)*, pages 3–12.

Allan, J., Wade, C., and Bolivar, A. (2003). Retrieval and novelty detection at the sentence level. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, SIGIR '03, pages 314–321, New York, NY, USA. ACM.

Aman, S. and Szpakowicz, S. (2007). Identifying expressions of emotion in text. pages 196–205.

Androutsopoulos, I., Koutsias, J., Chandrinos, K. V., Paliouras, G., and Spyropoulos, C. D. (2000). An evaluation of Naive Bayesian anti-spam filtering. In *Workshop on Machine Learning in the New Information Age*, pages 9–17.

Aron, A. and Aron, E. (1999). *Statistics for psychology*. Prentice Hall.

Asur, S. and Huberman, B. (2010). Predicting the future with social media. In *2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, pages 492–499. IEEE.

Baird, S. (2011). Twitter Celebrates 5 Years and 200 Million Users. `http://www.aolnews.com/2011/03/21/twitter-celebrates-5-years-and-200-million-users/`. [Online; accessed 21-April-2011].

Belkin, N. and Croft, W. (1987). Retrieval techniques. *Annual review of information science and technology*, 22:109–145.

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., Conway, M., McInerney, L., O'Hare, N., and Smeaton, A. (2009). Combining social network analysis and sentiment analysis to explore the potential for online radicalisation. In *Proceedings of the 2009 International Conference on Advances in Social Network Analysis and Mining-Volume 00*, pages 231–236. IEEE Computer Society.

Bermingham, A., Smeaton, A., Foster, J., and Hogan, D. (2008). DCU at the TREC 2008 Blog Track. In *The Seventeenth Text REtrieval Conference (TREC 2008) Proceedings. NIST*. Citeseer.

Bermingham, A. and Smeaton, A. F. (2009). A study of inter-annotator agreement for opinion retrieval. In *SIGIR '09: Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 784–785. ACM.

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.

Bethard, S., Yu, H., Thornton, A., Hatzivassiloglou, V., and Jurafsky, D. (2004). Automatic extraction of opinion propositions and their holders. In *2004 AAAI Spring Symposium on Exploring Attitude and Affect in Text*, page 2224.

Bollen, J., Mao, H., and Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*.

Bollen, J., Pepe, A., and Mao, H. (2009). Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. In *Proceedings of WWW 2009 Conference*.

Brew, A., Greene, D., and Cunningham, P. (2010a). Taking the pulse of the web: Assessing sentiment on topics in online media. In *Web Science Conference (WebSci 2010) at WWW 2010, Raleigh, North Carolina*.

Brew, A., Greene, D., and Cunningham, P. (2010b). Using crowdsourcing and active learning to track sentiment in online media. In *Proceeding of the 2010 conference on ECAI 2010: 19th European Conference on Artificial Intelligence*, pages 145–150. IOS Press.

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

Cambria, E., Speer, R., Havasi, C., and Hussain, A. (2010). Senticnet: A publicly available semantic resource for opinion mining. In *Proceedings of the 2010 AAAI Fall Symposium Series on Commonsense Knowledge*.

Carlson, N. (2011). Goldman to clients: Facebook has 600 million users. `http://www.msnbc.msn.com/id/40929239/ns/technology_and_science-tech_and_gadgets/`. [Online; accessed 21-April-2011].

Carvalho, P., Sarmento, L., Silva, M. J., and de Oliveira, E. (2009). Clues for detecting irony in user-generated contents: oh...!! it's "so easy" ;-). In *TSA '09: Proceeding of the 1st international CIKM workshop on Topic-sentiment Analysis for Mass Opinion*, pages 53–56, New York, NY, USA. ACM.

Choudhury, M., Saraf, R., Jain, V., Mukherjee, A., Sarkar, S., and Basu, A. (2007). Investigation and modeling of the structure of texting language. *International Journal on Document Analysis and Recognition*, 10(3):157–174.

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

Clarke, C. L., Kolla, M., Cormack, G. V., Vechtomova, O., Ashkan, A., Büttcher, S., and MacKinnon, I. (2008). Novelty and diversity in information retrieval evaluation. pages 659–666.

Cleverdon, C. (1967). The cranfield tests on index language devices. In *Aslib proceedings*, volume 19, pages 173–194. MCB UP Ltd.

Cormack, G. (2007). Email spam filtering: A systematic review. *Foundations and Trends in Information Retrieval*, 1(4):335–455.

Croft, B., Metzler, D., and Strohman, T. (2009). *Search Engines: Information Retrieval in Practice*. Addison-Wesley Publishing Company, USA, 1st edition.

Cunningham, P., Nowlan, N., Delany, S., and Haahr, M. (2003). A case-based approach to spam filtering that can track concept drift. In *The ICCBR*, volume 3.

Dan, O., Feng, J., and Davison, B. (2011). *Filtering microblogging messages for social TV*. WWW '11. ACM, New York, NY, USA.

Dang, H. T. (2008). Overview of the TAC 2008 opinion question answering and summarization tasks. In *Proceedings of Text Analysis Conference (TAC) 2008*.

Dave, K., Lawrence, S., and Pennock, D. M. (2003). Mining the peanut gallery: opinion extraction and semantic classification of product reviews. In *Proceedings of the 12th international conference on World Wide Web*, WWW '03, pages 519–528, New York, NY, USA. ACM.

Dawid, A. and Skene, A. (1979). Maximum likelihood estimation of observer error-rates using the em algorithm. *Applied Statistics*, pages 20–28.

Diakopoulos, N. A. and Shamma, D. A. (2010). Characterizing debate performance via aggregated Twitter sentiment. In *Conference on Human Factors in Computing Systems (CHI 2010)*.

Efron, M. (2010). Hashtag retrieval in a microblogging environment. In *SIGIR '10: Proceedings of the 33rd Annual ACM Conference on Research and Development in Information Retrieval*, pages 787–788, New York, NY, USA. ACM.

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

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.

Ekman, P. (1989). The argument and evidence about universals in facial expressions of emotion. *Handbook of social psychophysiology*, 143:164.

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

Esuli, A. and Sebastiani, F. (2006). SENTIWORDNET: A publicly available lexical resource for opinion mining. In *Proceedings of the 5th Conference on Language Resources and Evaluation (LREC-06)*, pages 417–422.

Fellbaum, C. (1998). *WordNet: An Electronical Lexical Database.* The MIT Press, Cambridge, MA.

Foster, J., Özlem Çetinoglu, Wagner, J., Roux, J. L., Hogan, S., Nivre, J., Hogan, D., and van Genabith, J. (2011). # hardtoparse: POS tagging and parsing the Twitterverse. In *Proceedings of AAAI-11 Workshop on Analysing Microtext*, San Francisco, CA.

Foster, J., Wagner, J., and van Genabith, J. (2008). Adapting a WSJ-trained parser to grammatically noisy text. In *Proceedings of ACL-08: HLT, Short Papers*, pages 221–224, Columbus, Ohio. Association for Computational Linguistics.

Freund, Y. and Schapire, R. E. (1996). Experiments with a New Boosting Algorithm. In *International Conference on Machine Learning*, pages 148–156.

Gabrilovich, E. and Markovitch, S. (2005). Feature generation for text categorization using world knowledge. In *International Joint Conference on Artificial Intelligence*, volume 19, page 1048. Citeseer.

Gamon, M. (2004). Sentiment classification on customer feedback data: Noisy data, large feature vectors, and the role of linguistic analysis. In *COLING '04: Proceedings of the*

20th International Conference on Computational Linguistics, page 841, Morristown, NJ, USA. Association for Computational Linguistics.

Garrett, S. (2010). Big Goals, Big Game, Big Records. `http://blog.twitter.com/2010/06/big-goals-big-game-big-records.html`. [Online; accessed 07-July-2011].

Gaughan, G. and Smeaton, A. (2005). Finding new news: Novelty detection in broadcast news. *Information Retrieval Technology*, pages 583–588.

Gerani, S., Carman, M., and Crestani, F. (2009). Investigating learning approaches for blog post opinion retrieval. *Advances in Information Retrieval*, pages 313–324.

Han, B. and Baldwin, T. (2011). *Lexical normalisation of short text messages: makn sens a #twitter*. HLT '11. Association for Computational Linguistics, Stroudsburg, PA, USA.

Hannon, J., McCarthy, K., and Smyth, B. (2011). Finding useful users on twitter: twittomender the followee recommender. *Advances in Information Retrieval*, pages 784–787.

Harman, D. (1995). *Overview of the third Text REtrieval Conference (TREC-3)*. DIANE Publishing.

Hayes, A. F. and Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. In *Communication Methods and Measures*.

He, B., Macdonald, C., He, J., and Ounis, I. (2008). An effective statistical approach to blog post opinion retrieval. In *Proceeding of the 17th ACM Conference on Information and Knowledge Management*, pages 1063–1072. ACM.

Healy, M., Delany, S., and Zamolotskikh, A. (2005). An assessment of case-based reasoning for short text message classification. In Creaney, N., editor, *Procs. of 16th Irish Conference on Artificial Intelligence and Cognitive Science, (AICS-05)*, pages 257–266.

Hearst, M. (2009). *Search User Interfaces*. Cambridge University Press.

Herring, S. C., Scheidt, L. A., Bonus, S., and Wright, E. (2004). Bridging the gap: A genre analysis of weblogs. In *HICSS '04: Proceedings of the Proceedings of the 37th*

Annual Hawaii International Conference on System Sciences (HICSS'04) - Track 4, page 40101.2, Washington, DC, USA. IEEE Computer Society.

Hu, M. and Liu, B. (2004). Mining opinion features in customer reviews. In *Proceedings of the National Conference on Artificial Intelligence*, pages 755–760. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999.

Hu, M. and Liu, B. (2006). Opinion extraction and summarization on the web. In *Proceedings Of The National Conference On Artificial Intelligence*, volume 21, page 1621. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999.

Jansen, B., Zhang, M., Sobel, K., and Chowdury, A. (2009). Twitter power: Tweets as electronic word of mouth. *Journal of the American Society for Information Science and Technology*.

Järvelin, K. and Kekäläinen, J. (2000). IR evaluation methods for retrieving highly relevant documents. pages 41–48.

Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. *Machine Learning: ECML-98*, pages 137–142.

Johansson, R. and Moschitti, A. (2010). Syntactic and semantic structure for opinion expression detection. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*, pages 67–76. Association for Computational Linguistics.

Karlgren, J., Eriksson, G., Sahlgren, M., and Täckström, O. (2010). Between bags and trees–constructional patterns in text used for attitude identification. *Advances in Information Retrieval*, pages 38–49.

Kelly, F. (2011). Attitudes in the tubes: An Irish site mines Twitter for political trends. `http://www.niemanlab.org/2011/02/attitudes-in-the-tubes-an-irish-site-mines-twitter-for-political-trends/`. [Online; accessed 27-August-2011].

Kennedy, J. (2011). Attitudes in the tubes: An Irish site mines Twitter for political trends. `http://www.siliconrepublic.com/new-media/item/20548-researcher-analyses-electio`. [Online; accessed 27-August-2011].

Kim, S.-M. and Hovy, E. (2007). Crystal: Analyzing predictive opinions on the web. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*.

Ku, L., Liang, Y., and Chen, H. (2006). Opinion extraction, summarization and tracking in news and blog corpora. In *Proceedings of AAAI-2006 Spring Symposium on Computational Approaches to Analyzing Weblogs*, pages 100–107.

Laboreiro, G., Sarmento, L., Teixeira, J., and Oliveira, E. (2010). Tokenizing microblogging messages using a text classification approach. In *AND '10: Proceedings of The Fourth Workshop on Analytics for Noisy Unstructured Text Data*, New York, NY, USA. ACM.

Liu, B. (2010). Sentiment analysis and subjectivity. *Handbook of Natural Language Processing Second Edition*, pages 1–38.

Liu, B., Hu, M., and Cheng, J. (2005). Opinion observer: analyzing and comparing opinions on the web. In *Proceedings of the 14th international conference on World Wide Web*, pages 342–351. ACM.

MacDonald, C. and Ounis, I. (2006). The TREC Blogs06 collection : Creating and analysing a blog test collection. Dept of Computing Science, University of Glasgow.

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*, pages 362–367.

Matsumoto, S., Takamura, H., and Okumura, M. (2005). 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*.

Metzler, D., Dumais, S., and Meek, C. (2007). Similarity measures for short segments of text. In *Advances in Information Retrieval*, Lecture Notes in Computer Science, chapter 5, pages 16–27.

Mishne, G. and de Rijke, M. (2006a). Moodviews: Tools for blog mood analysis. In *AAAI 2006 Spring Symposium on Computational Approaches to Analysing Weblogs (AAAI-CAAW 2006)*, pages 153–154. AAAI Press.

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

Mishne, G. and Glance, N. (2006). Predicting movie sales from blogger sentiment. In *AAAI 2006 Spring Symposium on Computational Approaches to Analysing Weblogs (AAAI-CAAW 2006)*.

Mishne, G. A. (2007). *Applied Text Analytics for Blogs*. PhD thesis, University of Amsterdam, Amsterdam.

Nakagawa, T., Inui, K., and Kurohashi, S. (2010). Dependency tree-based sentiment classification using crfs with hidden variables. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 786–794. Association for Computational Linguistics.

O'Connor, B., Balasubramanyan, R., Routledge, B. R., and Smith, N. A. (2010a). From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series. In *Proceedings of the International AAAI Conference on Weblogs and Social Media*.

O'Connor, B., Krieger, M., and Ahn, D. (2010b). TweetMotif: Exploratory search and topic summarization for Twitter. In *Proceedings of the Fourth International Conference on Weblogs and Social Media, ICWSM 2010, Washington, DC, USA, May 23-26, 2010*.

O'Hare, N., Davy, M., Bermingham, A., Ferguson, P., Sheridan, P., Gurrin, C., and Smeaton, A. F. (2009). Topic-dependent sentiment analysis of financial blogs. In *TSA 2009 - 1st International CIKM Workshop on Topic-Sentiment Analysis for Mass Opinion Measurement*, Hong Kong, China.

Ounis, de Rijke, M., Macdonald, C., Mishne, G., and Soboroff (2006). Overview of the TREC-2006 Blog Track. In *Proceeddings of the 15th Text REtrieval Conference (TREC 2006)*.

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)*.

Pang, B. and Lee, L. (2004). 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. Association for Computational Linguistics.

Pang, B. and Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135.

Pang, B., Lee, L., and Vaithyanathan, S. (2002). Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 79–86. Association for Computational Linguistics.

Phelan, O., McCarthy, K., Bennett, M., and Smyth, B. (2011). Terms of a feather: content-based news recommendation and discovery using Twitter. In *Proceedings of the 33rd European conference on Advances in information retrieval*, ECIR'11, pages 448–459, Berlin, Heidelberg. Springer-Verlag.

Pingdom (2010). Twitter, now 2 billion tweets per month. `http://royal.pingdom.com/2010/06/08/twitter-now-2-billion-tweets-per-month/`. [Online; accessed 26-April-2011].

Raykar, V., Yu, S., Zhao, L., Valadez, G., Florin, C., Bogoni, L., and Moy, L. (2010). Learning from crowds. *The Journal of Machine Learning Research*, 99:1297–1322.

Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P., and Riedl, J. (1994). Grouplens: an open architecture for collaborative filtering of netnews. In *Proceedings of the 1994 ACM conference on Computer supported cooperative work*, pages 175–186. ACM.

Riloff, E. and Wiebe, J. (2003). Learning extraction patterns for subjective expressions. In *Proceedings of the 2003 conference on Empirical methods in natural language processing-Volume 10*, pages 105–112. Association for Computational Linguistics.

Riloff, E., Wiebe, J., and Phillips, W. (2005). Exploiting subjectivity classification to improve information extraction. In *Proceedings of the National Conference On Artificial Intelligence*, volume 20, page 1106. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999.

Riloff, E., Wiebe, J., and Wilson, T. (2003). Learning subjective nouns using extraction pattern bootstrapping. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, pages 25–32. Association for Computational Linguistics.

Robertson, S. and Hancock-Beaulieu, M. (1992). On the evaluation of IR systems. *Information Processing & Management*, 28(4):457–466.

Robertson, S. E. and Hull, D. A. (2000). The TREC-9 Filtering Track final report. In *Proceedings of the 9th Text REtrieval Conference (TREC 2000)*.

Sakaki, T., Okazaki, M., and Matsuo, Y. (2010). Earthquake shakes Twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World Wide Web*, pages 851–860. ACM.

Salton, G. (1968). *Automatic Information Organization and Retrieval*. McGraw-Hill New York.

Saracevic, T. (2007a). Relevance: A review of the literature and a framework for thinking on the notion in information science. Part II: nature and manifestations of relevance. *Journal of the American Society for Information Science and Technology*, 58(13):1915 – 1933.

Saracevic, T. (2007b). Relevance: A review of the literature and a framework for thinking on the notion in information science. Part III: Behavior and effects of relevance.

Schütze, H., Hull, D., and Pedersen, J. (1995). A comparison of classifiers and document representations for the routing problem. In *Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 229–237. ACM.

Seki, Y., Evans, D., Ku, L., Chen, H., Kando, N., and Lin, C. (2007). Overview of opinion analysis pilot task at NTCIR-6. pages 265–278.

Seki, Y., Evans, D., Ku, L., Sun, L., Chen, H., Kando, N., and Lin, C. (2008). Overview of multilingual opinion analysis task at NTCIR-7.

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

Shamma, D. A., Kennedy, L., and Churchill, E. F. (2009). Tweet the debates: understanding community annotation of uncollected sources. In *Proceedings of the first SIGMM workshop on Social media*, WSM '09, pages 3–10, New York, NY, USA. ACM.

Shanahan, J. G., Qu, Y., and Wiebe, J., editors (2006). *Computing Attitude and Affect in Text: Theory and Applications*. Number 20 in the Information Retrieval Series. Springer.

Singhal, A. (2010). Being bad to your customers is bad for business. `http://googleblog.blogspot.com/2010/12/being-bad-to-your-customers-is-bad-for.html`. [Online; accessed 21-April-2011].

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*, pages 841–842. ACM.

Stoyanov, V. and Cardie, C. (2006). Toward opinion summarization: Linking the sources. In *Proceedings of the Workshop on Sentiment and Subjectivity in Text*, pages 9–14. Association for Computational Linguistics.

Su, X. and Khoshgoftaar, T. (2009). A survey of collaborative filtering techniques. *Advances in Artificial Intelligence*, 2009:4.

Tagliamonte, S. A. and Denis, D. (2008). LINGUISTIC RUIN? LOL! INSTANT MESSAGING AND TEEN LANGUAGE. *American Speech*, 83(1):3–34.

Takamura, H., Yokono, H., and Okumura, M. (2011). Summarizing a document stream. *Advances in Information Retrieval*, pages 177–188.

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*, pages 35–44. ACM.

Tong, S. and Koller, D. (2002). Support vector machine active learning with applications to text classification. *The Journal of Machine Learning Research*, 2:45–66.

Toutanova, K. and Manning, C. (2000). Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In *Proceedings of the 2000 Joint SIGDAT conference on Empirical methods in natural language processing and very large corpora: held in conjunction with the 38th Annual Meeting of the Association for Computational Linguistics-Volume 13*, pages 63–70.

Tumasjan, A., Sprenger, T., Sandner, P., and Welpe, I. (2010). Predicting elections with Twitter: What 140 characters reveal about political sentiment. In *International AAAI Conference on Weblogs and Social Media, Washington, DC*.

Weerkamp, W. and de Rijke, M. (2008). Credibility improves topical blog post retrieval. In *Proceedings of ACL-08: HLT*, page 923931, Columbus, Ohio. Association for Computational Linguistics, Association for Computational Linguistics.

Whitelaw, C., Garg, N., and Argamon, S. (2005). Using appraisal groups for sentiment analysis. In *Proceedings of the 14th ACM international conference on Information and knowledge management*, pages 625–631. ACM.

Wiebe, J. (1990). Identifying subjective characters in narrative. pages 401–406.

Wiebe, J. (1994). Tracking point of view in narrative. *Computational Linguistics*, 20(2):233–287.

Wiebe, J. and Bruce, R. (2001). Probabilistic classifiers for tracking point of view. *Progress in Communication Sciences*, pages 125–142.

Wiegand, M. and Klakow, D. (2010). Convolution kernels for opinion holder extraction. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 795–803. Association for Computational Linguistics.

Wilson, T., Wiebe, J., and Hoffmann, P. (2005). Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the 2005 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 347–354.

Wilson, T., Wiebe, J., and Hwa, R. (2004). Just how mad are you? Finding strong and weak opinion clauses. In *Proceedings of the National Conference on Artificial Intelligence*, pages 761–769. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999.

Wu, Y., Zhang, Q., Huang, X., and Wu, L. (2009). Phrase dependency parsing for opinion mining. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3-Volume 3*, pages 1533–1541. Association for Computational Linguistics.

Wunsch-Vincent, S. and Vickery, G. (2007). Participative web: User-created content. *Working Party on the Information Economy*, (2006):74.

Yang, Y., Zhang, J., Carbonell, J., and Jin, C. (2002). Topic-conditioned novelty detection. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 688–693. ACM.

Zhang, L., Zhu, J., and Yao, T. (2004). An evaluation of statistical spam filtering techniques. *ACM Transactions on Asian Language Information Processing (TALIP)*, 3(4):243–269.

Zhang, W., Yu, C., and Meng, W. (2007). Opinion retrieval from blogs. In *Proceedings of the sixteenth ACM conference on Conference on Information and Knowledge Management*, pages 831–840. ACM.

Zhang, X., Fuehres, H., and Gloor, P. A. (2010). Predicting Stock Market Indicators

Through Twitter I hope it is not as bad as I fear. In *Collaborative Innovations Networks Conference (COINs)*.