# Searching Twitter: Separating the Tweet from the Chaff

**Jonathan Hurlock, Max L. Wilson**
Future Interaction Technology Lab
Department of Computer Science, Swansea University, Swansea, SA2 8PP, UK
{csjonhurlock, m.l.wilson}@swansea.ac.uk

**Abstract**

Within the millions of digital communications posted in online social networks, there is undoubtedly some valuable and useful information. Although a large portion of social media content is considered to be babble, research shows that people share useful links, provide recommendations to friends, answer questions, and solve problems. In this paper, we report on a qualitative investigation into the different factors that make tweets 'useful' and 'not useful' for a set of common search tasks. The investigation found 16 features that help make a tweet useful, noting that useful tweets often showed 2 or 3 of these features. 'Not useful' tweets, however, typically had only one of 17 clear and striking features. Further, we saw that these features can be weighted as according to different types of search tasks. Our results contribute a novel framework for extracting useful information from real-time streams of social-media content that will be used in the design of a future retrieval system.

## Introduction

The casual statuses, exchanges, and communications posted online in social networking sites like Twitter are widely considered to mainly contain mindless babble. In fact in 2009, Pear Analytics (2009) classified 40% of Twitter communications to be so, with another 37.55% being conversational. Since then, posts to twitter have increased 500%, to more than 90 million per day. Yet despite the consensus that so much content may be entirely useless outside of one's social circle, research has clearly shown that people, for example, ask questions of their social networks (Morris, Teevan, & Panovich, 2010). Similarly, analyses have shown that many people share valuable information through posts and links on Twitter (boyd, Golder and Lotan 2010, Java, et al. 2007). Leveraging these millions of social communications, however, is often limited to either displaying the most recent or most popular communications, neither of which may be actually useful. This research, however, focuses on how we can sift through these real-time communications to identify potentially valuable pieces of useful information.

In the following sections, we first describe related work on how social networking sites are currently used in search systems, and discuss relevant research on social use of Twitter. We continue the related work by describing a few prototype systems that have been created to support the analysis of social media content. We then describe a study of participant searchers identifying valuable information in online Twitter streams, and present the results of a grounded theory analysis of the tweets that were deemed 'useful' and 'not useful'. Our research contributes a novel set of weighted filtering-features for both useful and not-useful tweets, as well as identifying several design recommendations for future systems.



## Related Work

Twitter is now a pretty well known microblogging service, which was founded in 2006. The original aim of Twitter was to allow its users to share short text based messages via SMS and via a web interface. Thus 'tweets' were born; a tweet is a message with a maximum length of 140 characters (similar to SMS). Tweets are typically published as fully public communications, unless users expressly opt-out, making Twitter an immense resource of casual information exchanges. Twitter differentiates itself from many other social networks by allowing unidirectional connections between users, meaning that Twitter users can 'follow' other users. As Twitter has grown, novel language use and standards have emerged organically within tweets, such as the reference/mention of other users with @theirusername. Further, Chris Messina (@chrismessina) ported the use of hashtags from Internet Relay Chat (IRC) forums to identify topical concepts in tweets, e.g. #ICWSM. Tweets can also be 'Retweeted' to ones own followers. Tweets typically include additional metadata such as time, source, and location.

As the popularity of social networking sites has exploded in the last decade, so has the amount of research into them. Many researchers have to varying degrees looked into what people ask their social networks. Morris, Teevan, and Panovich (2010), used a survey methodology, and found the most popular questions were to do with recommendation or opinion, typically asking for subjective and experience-based insights. Efron and Wignet (2010), on the other hand, used manual and social-computation methods to analyse the styles of actual Q&A questions mined from Twitter.

People also ask questions on other social forums, such as Question and Answer sites like Yahoo Answers[1] and Facebook Questions[2]. A recent service called Quora[3] has tried to improve the quality of answers by maintaining and constraining the community involved. These sources, however, can quickly become out of date (as opposed to Twitter and services like Wikipedia that are considered to be up-to-date, as described below). Aardvark[4], marketed as a social search engine, has tried to mix the real-time benefit of services like Twitter with the style of Q&A sites like Yahoo Answers. Rather than passively allowing users to answer questions, Aardvark actively targets users who are deemed to be 'experts' in a given field, and tries to contact them as soon as a question arrives that they may be able to answer. Aardvark, as a social search engine, is primarily focused on searching for people, rather than content in social networks (Horowitz & Kamvar, 2010), Horowitz & Kamvar's work builds on the findings that people prefer if they can make judgments on the ability of others to provide answers (Golbeck and Hendler 2006).

One value that many people take from Twitter is its up-to-date immediacy of content during real time situations. Prior work has evaluated the way in which information was shared during major events like the recent Iranian elections (Gaffney, 2010), and natural disasters (Vieweg, Hughes, Starbird, & Palen, 2010). Many twitter services now thrive on delivering breaking news[5], and tweets are often involved in 'real-time search'[6]. Although twitter content is frequently searched and analysed for content, including sentiment (Cheong & Lee, 2009), very little research, if any, has focused on characterizing tweets that contain 'useful' information.

Following the consensus that social media systems can contain useful and interesting information, several systems have focused on how to identify and deliver interesting content to users. Zerozero88 (Chen, Nairn, Nelson, Bernstein, & Chi, 2010) used an extensive algorithmic network-sensitive recommender system for delivering tweets to users. Once users had rated a few tweets, the system would continue to deliver similar content via a twitter account. Conversely, using a lightweight human-recommendation approach, FeedMe (Bernstein, Marcus, Karger, & Miller, 2010) created an extension for Google Reader[7] that allowed users to recommend content to other users. The key researchers involved in both Zerozero88 and FeedMe also created an exploratory user interface for browsing twitter streams using a tag-cloud that was generated by keywords assigned by the Yahoo! BOSS API. Similarly, Golovchinsky and Pickens (2010) proposed an exploratory search user interface that allowed users to explore tweets in their social network by pivoting and filtering over authors, tweets, and timelines. Further indicating that taking control over social media content is considered important. LinkedIn have recently invested in a service called Signal[8], which lets you enrich your twitter stream with a faceted browser for social media content.

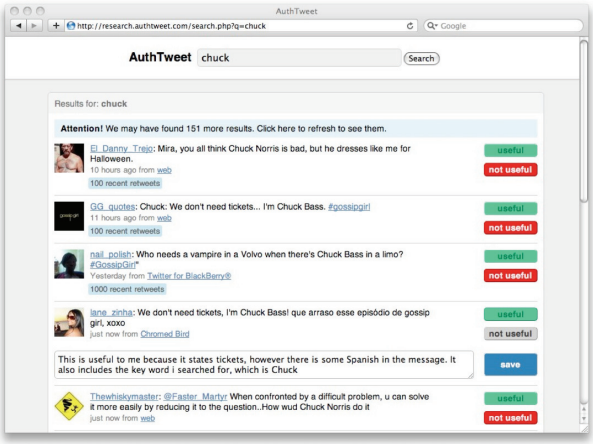


**Figure 1.** The search interface from the user study. Showing a user marking a tweet as a useful, and entering a reason.

The majority of the systems described above have been used to help users find popular content in their extended social networks. The work performed here has instead focused on the following research question: What makes a tweet 'useful' or 'not useful' when searching the global corpus of tweets. Our work intends to support those that use Twitter Search[9], or other search engines that return tweets, to find information. Further, however, characterizing useful tweets could also help the systems described above to rank and prioritize content to display. Consequently, our work is orthogonal to the design of exploratory user interfaces or recommender systems, but aims to identify useful content for all of them.

Recent work by Teevan, Ramage, and Morris (2011) has identified some common informational tasks that prompt twitter searches, including: temporally relevant information related to news or events, as well as social information regarding other users or popular trends on twitter. Below, we describe a study of the characteristics that are common to useful and useless tweets for these kinds of searches.

## User Study

The main hypothesis driving this work is that, although the main value of Twitter search systems have focused on up-to-date real time information, such online networks can be harvested for valuable and *relevant* information on a range of topics. Consequently, a user study was designed to help identify and elaborate on the types and sources of useful information found on Twitter. A custom-made search interface, shown in Figure 1, was provided that allowed users to provide both relevance feedback and qualitative

[1] http://answers.yahoo.com/
[2] http://www.facebook.com/questions/
[3] http://www.quora.com/
[4] http://vark.com
[5] http://twitter.com/breakingnews
[6] http://www.google.com/realtime
[7] http://www.google.com/reader/
[8] http://www.linkedin.com/signal/
[9] http://search.twitter.com/

comments for each tweet. The search would return tweets via the Twitter Search API[10]. Users were able to page through results, and new results were offered through an animated notification, as per Twitter's search. Using the custom search interface, users were able to select a tweet to be either 'useful' or 'not useful' via two buttons that were located next to each tweet. Once a user had selected a tweet as being 'useful' or 'not useful', an animation revealed an input textbox allowing them to leave a comment to explain their choice.

In the study, which lasted around an hour, participants began by providing demographic information, before completing 3 different types of informational search tasks. 10 minutes was provided for each search task. Following each search task, the relevance judgments and matching comments were discussed with the participant in a 5-minute semi-structured interview. A secondary review-based interface was generated to display the tweets they had rated and the comments they had made. The study concluded with a feedback questionnaire and final short debrief. On successful completion of the study, participants were entered into a draw to win one of three vouchers.

### Participants

20 staff and students were recruited from a range of departments across the university; 10 male and 10 female. Half of the participants were between 20-35 and the other between 36-50, in a roughly normal distribution. 90% of participants held a bachelor's degree or higher. When asked about their Internet usage, all 20 stated that they use the Internet everyday, with the majority spending more than 2 hours online. 12 stated they had used Twitter, 30% on a regular basis, and only two participants stated that they have attempted to search Twitter directly. We chose not to restrict our study to Twitter users, as we aim to find useful tweets that can be returned in any search system, including Google, which has an international user-base. Consequently, a mix of familiarity with Twitter allowed for a broader perspective on what constitutes useful social media content.

### Tasks

Three different types of search tasks were created: 1) a temporal monitoring task, 2) a subjective choice task, and 3) a location-sensitive planning task. These tasks are examples of tasks commonly performed over social media (Morris, Teevan, & Panovich, 2010). During the study, task order was counterbalanced in order to remove ordering effects. The temporal monitoring task involved tracking a progress of a current event. The most significant culturally relevant event at the time of the study was the BBC Proms[11]. Users were asked to identify interesting information about the on-going event.

In the subjective task, people were asked to find information that might help them decide whether to buy the new iPhone. Participants were asked to identify information that might help them to make the decision. In the location-sensitive planning task, participants were asked to find somewhere nice to eat lunch in London. Participants were asked to identify information that helped them decide where they might go for lunch.

## Analysis and Results

In total, participants rated 496 tweets, of which 482 were unique. Of the original 496 ratings, 52% were considered to be useful to participants. After splitting the data from the tasks into two sets ('useful' and 'not useful'), we used an inductive Grounded Theory (Glaser & Strauss, 2009) approach to reveal commonalities in the comments made about tweets. Grounded Theory is an established systematic procedure for identifying common topics and themes in qualitative text. Pieces of text are given 'codes' that represent their meaning. These codes are then grouped into themes, and used to produce underlying theories about the data. The rigor of our approach is detailed below.

Analysis began with both authors evaluating an initial set of 100 useful and 100 not-useful 'tweet+response' pairs independently. The authors then met and compared the codes created so far. This allowed us to both reflect on the dataset, and broaden our perspectives of the dataset and possible codes. We then continued to code the remaining tweets independently. We concluded the inductive coding by collating all the codes we had each created together and using a white-board affinity diagramming approach, commonly used to organise unstructured sets of ideas and concepts, to begin identifying relationships between themes and codes in the tweet+response pairs. We continued to discuss these codes and their definitions, using example tweet+response pairs, until the diagram stabilised. This process was then repeated for the not-useful tweets.

From the original set of more than 30 proposed codes, we settled on 16 codes for useful tweets, and were able to agree on 6 categories of 2-4 codes each. Similarly, we reduced the set of proposed not-useful codes and identified 17 codes that also fell into 6 categories of 2-5 each.

To validate these codes, Cohen's kappa (Cohen, 1960) was used to assess the inter-rater reliability of the authors. We achieved a high kappa score of 0.85 (Almost perfect agreement according to Landis and Koch (1977)) for the not-useful data set, as shown in Table 1. To further validate our results we introduced an independent untrained coder to the analysis. The independent judge was provided with a set of codes and definitions. Table 1 shows the Cohen scores achieved between all investigators, and that together

[10] http://dev.twitter.com/doc/get/search - we acknowledge that this already processes and selects tweets from the 'firehose' of content. We do not believe that his limited participants in finding useful and not-useful tweets during the study.

[11] The BBC Proms is an annual classical musical festival held in the UK. See: http://www.bbc.co.uk/proms/2010/

the three coders achieved a Fleiss's kappa (Fleiss, 1971) of 0.73 for the not-useful tweets.

At first, as shown in Table 1, we did not achieve such high scores with the 'useful' tweets, even between the two authors of the paper. Due to this difference in scores, we revisited the tweets and codes, discussed our findings, and sought to discover where the source of disagreement lay. From our investigations, we discovered that while the not-useful tweets typically had a single striking reason to be declared so, the useful tweets often had two or three valuable features. Author 1 observed an average of 2.14 codes per tweet-response pair in the useful tweets data set, with a range of 4 (Max: 5; Min: 1; STD: 0.90). Author 2 observed an average of 2.34 codes per tweet+response pair, with a range of 3 (Max: 4; Min: 1; STD: 0.92). For tweets deemed to be not useful, we saw a much lower average of 1.18, with a range of 2 (Max: 3; Min: 1; STD: 0.44).

As both Cohen and Fleiss analyses are performed when a single code is applied to a piece of text, we had originally asked the coders to choose 'the most appropriate code' for the tweet+response pair. Table 2 shows how the investigators easily applied different codes to the same tweet+response pair. Consequently, we sought to evaluate our codes using an analysis method that was suitable for multi-coding individual tweets. We performed a multi-coder, multi-coded kappa analysis detailed by Harris and Burke (2005), and achieved a score of 0.73 between the two coders, which is strong 'Substantial Agreement' according to Landis, and Koch (1977). This high score suggests that our original use of Cohen's kappa was indeed inappropriate. With our independent untrained validating judge, we also achieved multi-coded kappa score of 0.62; still a 'Substantial Agreement', and good given the high variability associated with multiple coding.

## Results

Tables 3 and 4 give us an overview of the codes, grouped by category, that were derived from the data, for both the useful and not-useful collections, respectively.

We saw four key reasons where the *content* of the tweet was directly useful. Some contained facts (e.g. times or prices) or increasingly common knowledge (e.g. problems with the iPhone). Others contained direct recommendations, or relayed insights from personal experiences. We also saw two types of tweets that the user found to be amenable, ones that were funny and ones that shared the searcher's perspective (e.g. Apple products are good or bad). We also saw two codes that focused on whether tweets were geographically or temporally still relevant (e.g. tweets in British prices). We also saw a key theme of trust, where users reported approving of trusted twitter accounts and recognising trustable avatars for those accounts. Also, links to authoritative or trustworthy websites were frequently recognised. Other links were also important, whether they provided more detailed information, rich media, or services (e.g. buying tickets).

There were also five key reasons that the content of tweets was not useful for the searcher. First tweets were frequently vague or introspective (for the author), or were quite directly not relevant by topic. While some tweets showed potential, it was easy for tweets to be too technical for the reader (containing jargon) or to contain errors (e.g. malformed URLs). There were 3 other reasons for tweets to be badly constructed: containing dead links, spam-style content, and being in a foreign language.

Like it was important for tweets to be temporally and geographically relevant, many tweets were deemed as not-useful because they were not current and about irrelevant locations. Similarly, Non-Trust was an issue, where users

| **Useful Tweets Data Set** | | | | |
|---|---|---|---|---|
| | Author 1 | Author 2 | Independent Coder | Fleiss' Kappa |
| Author 1 | - | 0.5065 | 0.5097 | |
| Author 2 | 0.5065 | - | 0.4607 | 0.4868 |
| Independent Coder | 0.5097 | 0.4607 | - | |
| **Not Useful Tweets Data Set** | | | | |
| | Author 1 | Author 2 | Independent Coder | Fleiss' Kappa |
| Author 1 | - | 0.8585 | 0.659 | |
| Author 2 | 0.8585 | - | 0.6856 | 0.7331 |
| Independent Coder | 0.659 | 0.6856 | - | |

**Table 1.** Showing Cohen's Kappa scores between multiple coders for both the Useful and Not Useful data sets. Also included is the Fleiss' Kappa score for each data set for the agreement between all three coders.

| Original Tweet Content | Reason given for being useful by participant | Code selected by author 1 | Code selected by author 2 |
|---|---|---|---|
| Offer - Caravaggio Restaurant book a table and EC3A 4DP - 2 course fixed price menu for £16.50 http://bit.ly/c7dl5w | specify the price of a 2 course meal - has location, but no reviews on the link | Location (Relevant) | Specific Information (In Tweet Content) |

**Table 2.** Showing how multiple codes could apply to one tweet and response pair.

were not happy with some pieces of information coming from non-authorities, and being linked to dubious websites. Further, not-useful tweets were often repeated content, or part of a conversation that would only be useful as a whole.

There were also three more subjective factors of not-useful tweets, including users disagreeing with the tweets (e.g. being pro or anti Apple), or not finding them funny.

## Analysis by Task

Tables 3 and 4 include counts for how frequently each code was applied to tweet+response pairs for each task.

*Temporal Search.* For the first task, useful and trusted links along with specific information, played main factors in deciding if a tweet was useful for that task. We also saw how other types of links, including media, were also frequent for the first task. The increased popularity of the media link code may have been influenced by the broadcast of the BBC Proms over the Internet. Media links, did not account for other tweets being regarded as useful for other tasks.

*Subjective Search.* For the subjective task, we were able to observe that experience with or of the subject matter was important to the information seekers. We also see two very interesting codes appear in this task, which are able to compliment each other, the first being shared sentiment, and secondly entertaining. Both of these codes are subjective in nature, which could be expected a subjective task. Useful links and experience were also played an important role in this task. Many participants found this task frustrating due to the amount of non-useful tweets; many of them were marked as SPAM or untrustworthy.

*Location Sensitive Search.* In the third (location-sensitive) task, we again see a high dependency on specific and useful information. However for this task, specific information played a more important role. As suspected we also see location sensitivity as an important factor, dominating this task with 85% of reasons to why location sensitivity is useful being allocated to this task. In this task, we see that trust, in the form of avatars and authors played an important role, with 2 tweet+response pairs being coded as useful because of the participant trusting the avatar, and a further 6 being coded as trusted author. Further, we see the introduction of direct recommendation and experience playing a part in why a participant found a tweet useful. Perhaps indicating a need for knowledge of first hand experience from someone who has been to a lunch venue in London, rather than a commercial entity trying to sell an experience or product.

## Relevance Judgments for Tasks

In the post-task interviews, we asked users to informally augment their relevance judgments with scores out of 5. Overall, the mean score for all rated tweets over all three tasks was 2.2, indicating a very low relevancy score. Individually, the first task, which was temporal in nature, scored 2.7. The second task, which involved users search for information regarding purchasing an iPhone, scored a very low 1.25. The third and final task, which was a

| In Tweet Content | | T1 | T2 | T3 |
|---|---|---|---|---|
| **Experience** | Someone reporting a personal experience, but not necessarily suggestion / direction. | 15 | 12 | 13 |
| **Direct Recommendation** | Someone making a direct recommendation, but not necessarily relaying a personal experience. | 3 | 3 | 20 |
| **Social Knowledge** | Containing information that is spreading socially, or becoming general knowledge. | 7 | 6 | 6 |
| **Specific Information** | Where facts are listed directly in tweets e.g. prices, times etc. | 51 | 10 | 47 |
| Reflection on Tweet | | | | |
| **Entertaining** | The reader finds them amusing. | 1 | 3 | 2 |
| **Shared Sentiment** | The reader agrees with the author of the tweet. | 1 | 2 | 1 |
| Relevant | | | | |
| **Time** | The time is current. | 14 | 0 | 2 |
| **Location** | The location is relevant to the query. | 6 | 1 | 40 |
| Trust | | | | |
| **Trusted Author** | The twitter account has a reputation / following. | 3 | 2 | 6 |
| **Trusted Avatar** | The visual appearance cultivates trust. | 2 | 0 | 2 |
| **Trusted Link** | A link to a trustworthy recognizable domain. | 14 | 1 | 7 |
| Links | | | | |
| **Actionable Link** | The user can perform a transaction by using the link (heavily dependent on trust). | 9 | 0 | 0 |
| **Media Link** | The link is to rich multimedia content. | 9 | 0 | 0 |
| **Useful Link** | The link provides valuable information content, e.g. authoritative information, educated reviews, and discussions. | 61 | 30 | 43 |
| Meta Tweet | | | | |
| **Retweeted Lots** | Its information that others have passed on lots. | 4 | 0 | 4 |
| **Conversation** | It is part of a series of tweets, and they all need to be useful. | 1 | 4 | 4 |

**Table 3.** The 16 codes and the 6 categories extracted from responses and tweet pairs from the useful tweets. Further, columns 3-5 show how frequently each was associated with the temporal (T1), subjective (T2) and location-sensitive (T3) tasks.

| Tweet Content | | T1 | T2 | T3 |
|---|---|---|---|---|
| **No Information** | Absence of anything, event factual points. | 16 | 14 | 12 |
| **Introspective** | Personal content and personal thoughts for no social benefit. | 5 | 5 | 8 |
| **Off Topic** | Result not related to the query given / TF-IDF irrelevant | 27 | 21 | 18 |
| **Too Technical** | The content requires specific domain knowledge the reader doesn't possess. | 1 | 2 | 2 |
| **Poorly Constructed** | Tweets that may have grammatical/spelling errors, or malformed URLs. | 3 | 2 | 3 |
| Bad Tweets | | | | |
| **SPAM** | Irrelevant or inappropriate messages. | 0 | 17 | 2 |
| **Wrong Language** | Messages sent in a foreign language of that to the reader. | 3 | 2 | 1 |
| **Dead Link** | A URL which does not work i.e. 404 | 2 | 4 | 3 |
| Not Relevant | | | | |
| **Time** | Out of date content. | 0 | 1 | 1 |
| **Location** | Wrong geographic location. | 2 | 7 | 2 |
| Trust | | | | |
| **Un-trusted Author** | An author the reader feels at un-eased by or suspicious of. | 4 | 7 | 1 |
| **Un-trusted Link** | A link the reader feels is suspicious | 4 | 7 | 2 |
| Subjective | | | | |
| **Perspective Oriented** | A tweet that is perspective centric, meaning the author is providing their views or projecting an attitude on a subject matter or to a subject/reader. | 2 | 3 | 2 |
| **Disagree with Tweet** | A conflict of agreement between the reader and the author | 2 | 2 | 1 |
| **Not Funny** | A tweet that is aimed to be humorous, which the reader does not feel is humorous. | 1 | 1 | 1 |
| Meta Tweet | | | | |
| **QnA** | Part of a conversation, reader desires the whole conversation, not just the question or the answer, but both the question and answer | 2 | 4 | 9 |
| **Repeated** | Content the reader has seen before | 3 | 7 | 1 |

**Table 4.** The 17 codes, in 6 categories, extracted from responses and tweet pairs from the not-useful tweets. Further, columns 3-5 show how frequently each was associated with the temporal (T1), subjective (T2) and location-sensitive (T3) tasks.

location-sensitive task, averaged at 2.75. No participant rated a tweet with a score of 5 (very relevant). 20% of participants, however, gave a score of 0 (not relevant) during the second subjective task, but not in the temporal and location-sensitive tasks.

## Common Patterns

As well as statistical analysis of the codes we were able to pick up on structural traits of tweets. Some of the structures that we were able to extract combined several of our codes combined together to make a structure. One in particular, which we called a teaser, combined codes for specific information and a link, which accounted for 22% of the useful tweet+response pairs. Another 13% were coded holding both specific information & location codes, which we attributed mainly to the location-based task.

Another structural concept we came across was actually a code QnA which is where a user could only see part of a question whether that be the question itself or an answer to a question, but could not see both parts, or multiple answers. The QnA code was found in 6% of the not-useful dataset and highlights the need for returning responses to question-tweets returned by a search.

Twitter itself has tackled some of these concepts when browsing its website. For instance the embedding of images and some videos in its new layout. As well as the 'in reply to' feature shown when browsing the site (Williams, 2010). These features have failed to make it over to Twitter's search service.

## Future Work and Design Recommendations

The aim of our study was to identify the traits of tweets that provide useful information, and of course those that do not. The key design recommendation we can make, therefore, is that future search systems, like the one we are now developing, could use the 16 positive and 17 negative factors to identify valuable and *relevant* tweets. The majority of these features are objective and easily identifiable characteristics.

We also found additional evidence for identifying tweets from authors that people may recognize. In lieu of identifying tweets that are socially connected to the searcher, our analysis suggests that authority measures, such as TunkRank[12] and Klout[13], could also be used to assess estimated trustworthiness.

We were also surprised to see that some codes, such as 'Retweeted lots', did not feature as highly as we had expected. With just under half of participants stating they have not used Twitter, and only 30% stating they use it regularly, we suspect that unfamiliarity with Twitter specific features may be a reason. If we are to export

[12] http://www.tunkrank.com
[13] http://www.klout.com/

knowledge from Twitter to the masses, however, we must ask how do we best explain these features to users, in an intuitive and simple way to understand.

Although most of the post-task interviews simply elaborated on the points noted by participants during the study, a few additional factors were identified. One potentially interesting additional factor was the impact of a tweeter's avatar. Many users suggested that avatars were a factor in choosing whether a tweet was trustworthy or not; most stating that they like to see faces of individuals. Several participants stated that they thought they would be able to tell if a tweeter had similar preferences to them by just looking at their avatar. One participant, for example, said: "Why would a baby give me a free phone?? Automatically suspect a con or a virus!" This suggests that both the type and presence of an avatar have an affect on the trustworthiness of tweets. On discussing the importance of trust, another participant said "… Also think I know this tweeter - a friend of a friend - so might be inclined to try the restaurant anyway!" These findings about trust echo the principles of Aardvark's social network routing efforts, but the emphasis on visual avatar judgments is important to note for future systems.

When asked if users were able to guess where authors were when they tweeted, or when they tweeted, most participants stated they were not aware of these factors, unless some specifically said 'I am in…' It appeared, through discussions with participants, that metadata played a very small role in their search experience. This may be a factor of the way results are displayed in Twitter (and our customer search interface), but could imply that metadata is more useful for the algorithms than the searcher.

In regards to query size, participants also mentioned frustration when searching, noting that longer queries returned much fewer results, or no results at all. Users noted that shorter queries, using one or two general terms were much more productive. This is likely due to the short limited size of tweets. Social search user interfaces may wish to encourage shorter, more general queries, but will have to work harder to identify the implied contexts associated with them.

Finally, there are some items in our codes for useful tweets that are mirrored in our not-useful codes. Further exploration of this relationship would be both interesting and beneficial. Such an analysis would help measure the influence of different features on a single tweet, especially if it contains both useful and non-useful features.

### In-progress Social-media Retrieval System

The codes we have found are currently being built into a social-media information retrieval system. So far system is able to perform n-gram extraction, and location relevance by extracting location names, using part of speech tagging and proper noun extraction of n-grams. By passing this list of n-grams through a database we are able to check, for example, if any locations are mentioned. As well as this, the system is able to search for geographic clues via other language clues such as the mentioning of time, and currency. As well as looking at in-tweet content, we are able to extract a great detail of information via the Twitter API. We are able to, for example, grab the user's profile location, as well the geo-location of a single a tweet if tagged, and pass it through our system to extract the location of the user.

These analysis tools will be used to implement a set of ranking features, based on the 33 codes determined in this paper, to score tweets for 'usefulness'. We have begun to implement a modified locality sensitive hashing algorithm (LSH - Indyk and Motwani 1998), which will enable us to cluster tweets by content in near real time. A deviation of the LSH algorithm was used by Petrovic, Osborne and Lavrenko (2010) for performing first-story detection within Twitter. By taking advantage of filtering services, such as that offered by DataSift[14], we hope to perform near real time analysis of Tweets, whilst factoring in our 33 codes.

### Perspective-oriented Retrieval

Although the majority of our codes can be objectively identified, there were a few features that were subjective or perspective-oriented. One clear example was whether the searcher and the tweet-author were both pro or anti companies like Apple or Microsoft. Such perspective-oriented examples were clearly seen between codes 'Entertaining' (in tweet content from Table 3) and 'Not Funny' (subjective from Table 4). This poses a larger question: how could we tailor a search system to take into a person's emotions and their personal preferences? One of our larger research aims is to investigate this question. By allowing a user to choose if they want tweets that they may find agreeable or not, we could make their information seeking experience more insightful in ways they may have been ignorant to before.

## Conclusions

In this study, we have used a range of human searching tasks to identify the types of social-networking communications that are both useful and not-useful. We have made three clear contributions. First, we used an inductive grounding theory analysis to identify 16 factors of useful content and 17 factors of non-useful content. Second, our results highlight that useful tweets typically have two or three strengths, while not-useful tweets often have a single and clearly identifiable fault. Third, we have identified that these factors apply with different weightings to different types of common social-media search tasks.

We are currently using these discovered factors to develop a social-media information search system, which will be used to cross-validate these weighted factors, across a larger range of tasks. The research reported here, however, contributes a novel weighted framework, for designing future social-media search systems, that can be used to extract *valuable and useful* information from

[14] http://www.datasift.net/

social-media communications, rather than simply the most recent or most popular.

## Acknowledgements

This work is part-funded by the European Social Fund (ESF) through the European Union's Convergence programme administered by the Welsh Assembly Government, and performed in collaboration with Pingar. Thanks also to our study participants and to Mathew Wilson for supporting the analysis phase of the project.

## References

Bernstein, M. S., Marcus, A., Karger, D. R., & Miller, R. C. (2010). Enhancing directed content sharing on the web. *CHI '10: Proceedings of the 28th international conference on Human factors in computing systems* (pp. 971-980). New York, NY, USA: ACM.

boyd, d., Golder, S., & Lotan, G. (2010). Tweet, Tweet, Retweet: Conversational Aspects of Retweeting on Twitter. *HICSS-43, IEEE: Proceedings of the 2010 43rd Hawaii International Conference on System Sciences. 43.* Washington, DC, USA: IEEE Computer Society.

Chen, J., Nairn, R., Nelson, L., Bernstein, M., & Chi, E. (2010). Short and tweet: experiments on recommending content from information streams. *CHI '10: Proceedings of the 28th international conference on Human factors in computing systems* (pp. 1185-1194). New York, NY, USA: ACM.

Cheong, M., & Lee, V. (2009). Integrating web-based intelligence retrieval and decision-making from the twitter trends knowledge base. *SWSM '09: Proceeding of the 2nd ACM workshop on Social web search and mining.* New York, NY, USA: ACM.

Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement , 20* (1), 37-46.

Efron, M., & Winget, M. (2010). Questions are content: a taxonomy of questions in a microblogging environment. *ASIS&T '10: Proceedings of the 73rd ASIS&T Annual Meeting on Navigating Streams in an Information Ecosystem - Volume 47. 47.* Silver Springs, MD, USA: American Society for Information Science.

FilmTrust: Movie recommendations using trust in web-based social networks. (2006). *Proceedings of the IEEE Consumer Communications and Networking Conference* (pp. 282-286). Piscataway, NJ, USA: IEEE Press.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin , 76* (5), 378-382 .

Gaffney, D. (2010). #iranElection: Quantifying Online Activism. *Proceedings of the WebSci10: Extending the Frontiers of Society On-Line.* Raleigh, NC, USA: Web Science Trust.

Glaser, B. G., & Strauss, A. L. (2009). *The Discovery of Grounded Theory: strategies for qualitative research.* Piscataway, New Jersey, USA: Transaction Publishers.

Golovchinsky, G., & Pickens, J. (2010). Interactive information seeking via selective application of contextual knowledge. *IIiX '10: Proceeding of the third symposium on Information interaction in context* (pp. 145-154). New York, NY, USA: ACM.

Harris, J. K., & Burke, R. C. (2005). Do you see what I see? An application of inter-coder reliability in qualitative analysis. *American Public Health Association 133rd Annual Meeting & Exposition.* Washington, DC, USA: American Public Health Association.

Horowitz, D., & Kamvar, S. D. (2010). The anatomy of a large-scale social search engine. *WWW '10 Proceedings of the 19th international conference on World wide web . 19*, pp. 431-440. New York, NY, USA: ACM.

Indyk, P., & Motwani, R. (1998). Approximate Nearest Neighbors: Towards Removing the Curse of Dimensionality. *STOC '98: Proceedings of the thirtieth annual ACM symposium on Theory of computing* (pp. 604-613). New York, NY, USA: ACM.

Java, A., Song, X., Finin, T., & Tseng, B. (2007). Why we twitter: understanding microblogging usage and communities. *WebKDD/SNA-KDD '07: Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis. 9*, pp. 56-65. New York, NY, USA: ACM.

Landis, R. J., & Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics , 33* (1), 159-174.

Morris, M. R., Teevan, J., & Panovich, K. (2010). What Do People Ask Their Social Networks, and Why? A Survey Study of Status Message Q&A Behavior. *CHI '10 Proceedings of the 28th international conference on Human factors in computing systems . 28*, pp. 1739-1748. New York, NY, USA: ACM.

Pear Analytics. (2009, August). *Twitter Study - 2009.* Retrieved from Pear Analytics: http://www.pearanalytics.com/blog/wp-content/uploads/2010/05/Twitter-Study-August-2009.pdf

Petrović, S., Osborne, M., & Lavrenko, V. (2010). Streaming first story detection with application to Twitter. *HTLT'10: Proceeding HLT '10 Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics* (pp. 181-189). Stoudsburg, PA, USA: ACL.

Teevan, J., Ramage, D., & Morris, M. R. (2011). #TwitterSearch: a comparison of microblog search and web search. *WSDM '11: Proceedings of the fourth ACM international conference on Web search and data mining* (pp. 35-44). New York, NY, USA: ACM.

Vieweg, S., Hughes, A. L., Starbird, K., & Palen, L. (2010). Microblogging during two natural hazard events: what twitter may contribute to situational awareness. *CHI '10: Proceedings of the 28th international conference on Human factors in computing systems.* New York, NY, USA: ACM.

Williams, E. (2010, September 14). *A Better Twitter.* Retrieved from Twitter Blog: http://blog.twitter.com/2010/09/better-twitter.html

Wilson, M. L., & Elsweiler, D. (2010). Casual-leisure Searching: the Exploratory Search scenarios that break our current models. *HCIR '10: 4th International Workshop on Human-Computer Interaction and Information Retrieval.* New York, NY, USA: ACM.

Yardi, S., Romero, D., & Schoenebeck, G. (2010). Detecting spam in a Twitter network. *First Monday , 15*, 2.